THE UNIVERSITY OF CHICAGO

CONTENT KNOWLEDGE FOR TEACHING AMONG MATHEMATICS TEACHERS: INVESTIGATING ITS INEQUITABLE DISTRIBUTION AND CAUSAL IMPACTS ON INSTRUCTIONAL QUALITY AND STUDENT OUTCOMES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE SOCIAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPARATIVE HUMAN DEVELOPMENT

BY

MENGYUAN LIANG

CHICAGO, ILLINOIS

AUGUST 2024

For my beloved parents,

Jiangli Meng and Rongtao Liang,

my grandparents,

Jiebin Meng, Xiaoying Liu, Guizhen Liang, Shuliang Han,

and my great grandma, “A-zu”, Yueying Bi.

I learned that courage was not the absence of fear, but the triumph over it. The brave is not the person who does not feel afraid, but the person who conquers that fear.

Nelson Mandela

Table of Contents

List of Tables viii
List of Figures x
Acknowledgments xi
Abstract xiii
Chapter One. Overview 1
I. Research Questions 3
II. Research Design 4
III. Measurement 5
IV. Analyses 6
Chapter Two. Inequitable Distribution of Teachers' Content Knowledge for Teaching Across Elementary and Secondary Math Classes 8
I. Introduction 8
II. Literature Review 11
1. Prior research on allocation of high-quality teachers 11
2. Measurement of teacher knowledge in mathematics 15
3. Hypotheses based on prior research 17
III. Data 18
1. Sample description 18
2. Measurement 19
3. Analytic strategy 20
IV. Results 22
1. Variance decomposition of teacher CKT 22
2. Systematic sorting of teacher CKT across schools 23
3. Systematic sorting of teacher CKT within schools 25
V. Conclusion and discussion 26
Chapter Three. Does Teacher Content Knowledge Impact the Quality of Instructional Practices? A Causal Analysis 29
I. Introduction 29
II. Literature Review 31
1. Measurement of instructional quality: using classroom observation and student perception 31
(1) Classroom observations 32
(2) Student perception 34
(3) Cross-instrument structure of instructional quality 36

2. Relationships between CKT and instructional quality 39
III. Hypotheses 41
IV. Data description 43
V. Analytic results 45
1. Factor analysis on classroom ratings 45
(1) Exploratory factor analysis 46
(2) Confirmatory factor analysis 50
2. Causal analysis of CKT effect on instructional quality 53
(1) Methodological framework 53
(2) Analytic strategy 56
(3) Analytic results of causal effects 59
VI. Conclusion 62
Chapter Four. Does Instructional Quality Mediate the Impacts of CKT on Student Achievement in Mathematics: Evidence from A Casual Mediation Analysis 66
I. Introduction 66
II. Literature Review 68
1. Evidence on relationship between CKT and learning outcomes. 68
2. Relation between instructional quality and student outcomes. 70
3. Mediation studies of CKT, instructional quality, and student learning 72
4. Summary of research gaps and potential contribution of this study 75
III. Hypotheses 76
IV. Data description 77
V. Methodology 78
1. A General framework for causal mediation analysis. 78
(1) Definition of the causal effect 78
(2) Identification assumptions 82
(3) Model specifications 85
Step 1. Analyzing the total effect of the treatment on the outcome conditioning on the covariates 88
Step 2. Analyzing the treatment effect on each mediator conditioning on the covariates 89
Step 3. Analyzing the mediated effect on the outcome via the mediator conditioning on the covariates 90
VI. Analytic results 92
1. Main variables in the causal mediation analysis. 92

Outcome variable. 92
Mediators. 92
Treatment variable. 93
Step 1 Results: Total effects of CKT on student achievement 93
Step 2 Results: Effect pathways of CKT on instructional quality 94
Step 3 Results: Evidence on indirect effects of CKT through changing instructional quality 95
Discussion of the analytic results 98
VII. Conclusion 100
Chapter Five. Summary and Future Directions 104
I. Overview of the Key Findings 104
Study 1: Inequitable Distribution of CKT: Evidence from Natural Variation of The Year 1 Baseline Observational Data 104
Study 2: Causal Relationship Between CKT and Instructional Quality 105
Study 3: Mediation of Instructional Quality in the Relationship Between CKT and Student Achievement 106
II. Implications for Policy and Practice 108
III. Limitations and Future Directions 109
Tables 111
Figures 141
Reference 158
Appendices 166
Alternative Model for Step 3. Analyzing the mediated effect on the outcome via the mediator and the treatment-by-mediator interaction conditioning on the covariates 166
Appendix Tables 168
Appendix Figure 186

List of Tables

Table 1. 1 Core Elements of the MET study by Year....111
Table 1. 2 Participation of the MET Study by Year, Study Type, and Unit Level 113
Table 2. 1 Descriptive statistics of participating teachers' characteristics 114
Table 2. 2 Associations of CKT latent scores with teacher characteristics 115
Table 2. 3 Descriptive statistics of covariates at teacher-level and school-level within segmented sub-samples 116
Table 2. 4 Model specifications for Hierarchical Generalized Linear Model 118
Table 2. 5 Natural variation by grade, by state and variance decomposition of full sample and subsamples of various levels of schools 119
Table 2. 6 Variance comparison by different combinations of fixed effects (Full sample) 121
Table 2. 7 Variance comparison of CKT latent scores across models including different covariates by school levels (With adjustment for district and grade fixed effects) 122
Table 2. 8 Inequality in CKT distribution by prior performance levels 123
Table 2. 9 Inequality in CKT distribution by Free/Reduced-priced Lunch Status 124
Table 2. 10 Inequality in CKT distribution by Minority Status 125
Table 2. 11 Inequality in CKT distribution by English Language Learner Status 126
Table 2. 12 Inequality in CKT distribution by Special Education Status 127
Table 3. 1 Pairwise correlations of MQI items 128
Table 3. 2 Factor loadings of MQI items 129
Table 3. 3 Descriptive statistics of SEM-constructed factors by grade levels 130
Table 3. 4 Pairwise correlations of SEM-constructed factors 131
Table 3. 5 Associations of CKT latent scores with teacher characteristics 132
Table 3. 6 Analytic results for CKT impacts on MQI 133
Table 3. 7 Analytic results for CKT impacts on MQI dimension scores 134
Table 4. 1 Descriptive statistics of main variables by grade levels 135
Table 4. 2 Analytic Results for Step 1: Treatment Effects on Student Achievement- Total Effects, α_{0200} 136
Table 4. 3 Analytic Results for Step 2: Treatment Effects on Mediators, γ_{0100} 137
Table 4. 4 Analytic Results for Step 3: Controlled Direct Treatment Effects, β_{0100} and Mediator Effects on Student Achievement given Treatment, β_{0200} 138
Table 4. 5 Indirect effects derived from multi-step regression estimates 140
Table A2. 1 Model Specifications for Weighted Regression 168
Table A2. 2 Natural variation by grade, by state and variance decomposition of full sample and subsamples of various levels of schools (Weighted regression) 169
Table A2. 3 Inequality in CKT distribution by prior performance levels (Weighted regression)170
Table A2. 4 Inequality in CKT distribution by Free/Reduced-priced Lunch Status (Weighted regression) 171
Table A2. 5 Inequality in CKT distribution by Minority Status (Weighted regression) 172
Table A2. 6 Inequality in CKT distribution by English Language Learner Status (Weighted regression) 173
Table A2. 7 Inequality in CKT distribution by Special Education Status (Weighted regression) 174
Table A3. 1 CLASS theoretical domains 175
Table A3. 2 MQI domains 176

Table A3. 3 Tripod items 177
Table A3. 4 Detailed Fit Indices for Elementary School Data 179
Table A3. 5 Detailed Fit Indices for Secondary School Data 180
Table A3. 6 Analytic results for CKT impacts on CLASS and Tripod composite scores 181
Table A3. 7 Analytic results for dimension scores (outcomes centered by randomization block) 182
Table A3. 8 Analytic results for CKT impacts on Tripod 7Cs in high school classrooms (outcomes centered by randomization block) 183
Table A4. 1 Analytic Results for Step 3: Treatment Effects on Outcomes with Mediator and Treatment-by-mediator Interaction 184

List of Figures

Figure 1. 1 Structure of dissertation project 141
Figure 3. 1 Diagrams indicating structures of three different SEMs for CLASS 142
Figure 3. 2 SEM Results of the final structure in path diagram (Elementary data) 144
Figure 3. 3 SEM Results of the final structure in path diagram (Secondary data) 145
Figure 4. 1 Pairwise relationship between CKT latent scores (group mean centered) and student math achievement (group mean centered) in elementary-school sample 146
Figure 4. 2 Pairwise relationships between CKT latent scores and MQI, and between MQI and student achievement in elementary-school sample, all variables group mean centered 147
Figure 4. 3 Pairwise relationships between CKT latent scores and CLASS composite scores, and between CLASS and student achievement in elementary-school sample, all variables group mean centered 148
Figure 4. 4 Pairwise relationships between CKT latent scores and Tripod composite scores, and between Tripod and student achievement in elementary-school sample, all variables group mean centered 149
Figure 4. 5 Pairwise relationship between CKT latent scores (group mean centered) and student math achievement (group mean centered) in middle-school sample 150
Figure 4. 6 Pairwise relationships between CKT latent scores and MQI, and between MQI and student achievement in middle-school sample, all variables group mean centered 151
Figure 4. 7 Pairwise relationships between CKT latent scores and CLASS composite scores, and between CLASS and student achievement in middle-school sample, all variables group mean centered 152
Figure 4. 8 Pairwise relationships between CKT latent scores and Tripod composite scores, and between Tripod and student achievement in middle-school sample, all variables group mean centered 153
Figure 4. 9 Pairwise relationship between CKT latent scores (group mean centered) and student math achievement (group mean centered) in high-school sample 154
Figure 4. 10 Pairwise relationships between CKT latent scores and MQI, and between MQI and student achievement in high -school sample, all variables group mean centered 155
Figure 4. 11 Pairwise relationships between CKT latent scores and CLASS composite scores, and between CLASS and student achievement in high -school sample, all variables group mean 156
Figure 4. 12 Pairwise relationships between CKT latent scores and Tripod composite scores, and between Tripod and student achievement in high -school sample, all variables group mean centered 157
Figure A4. 1 Effective Instructional Practices Identified by Brophy and Good 186

Acknowledgments

I am privileged and deeply grateful to have Dr. Guanglei Hong as my advisor. Her exceptional guidance and continuous support have been crucial in my intellectual growth. She consistently provided inspiration and advice throughout my academic journey and graciously forgave me for all the mistakes I made. I have gained invaluable knowledge from seven years of working with her. Her great dedication, passion, professionalism, high standards, and remarkable work ethic have set a lifelong example that I aspire to follow.

I would also like to express my gratitude to the other members of my committee. Dr. Stephen Raudenbush was immensely helpful during my dissertation journey, aiding me in organizing my thoughts and offering insightful and critical feedback that significantly contributed to the development and refinement of my analytic models. Dr. Robert Gibbons's outstanding work in measurement and item response theory have been a profound source of inspiration for me. His course on statistical application, enriched by his sense of humor and insights from navigating academic life, helped me truly enjoy statistics and taught me to approach methodological challenges with confidence and optimism. I am very honored to have had Dr. Benjamin Kelcey on my committee; his expertise on mediation analysis on content knowledge, instruction and student outcomes, and helpful feedback have been immeasurable.

I would always treasure my time in the University of Chicago where I received rigorous academic training and forged my character as a young adult, experiencing both tears and laughter, despair and hope. I sincerely appreciate all the people I met during my academic journey in the United States, including my dear friends, classmates, teachers, staff, neighbors and even kind strangers whose names will fill a whole page if all listed. Their warmth and kindness,

especially during the freezing-cold Chicago winters, have made Chicago feel like a second hometown to me.

At the end, I want to thank my family for their unconditional love and great support: my mother, Jiangli Meng, a wise and elegant woman who steadfastly believes that women are smart and strong enough to conquer any challenges in career and life; my father, Rongtao Liang, the guardian and soldier of our family, for raising me to be an independent woman whom he believes is capable of learning and mastering any knowledge and skills. I also extend my gratitude to my extended family in China, who have always supported us and formed a close-knit bond around us.

Abstract

This dissertation investigates the inequitable distribution and causal impacts of mathematics teachers' content knowledge for teaching (CKT) on instructional quality and student learning outcomes. Using the rich information in the Measures of Effective Teaching (MET) project longitudinal database, the study addresses three core research questions: the systematic inequality of CKT associated with student backgrounds, the causal impacts of CKT on the fine-grained aspects of instructional quality, and the mediating role of instructional quality in the relationship between CKT and student achievement.

The research comprises three interconnected studies. The first study uses latent scores of mathematical CKT as a direct measure for teacher knowledge and explores its distribution across different school levels. Unlike conventional qualification indicators such as advanced degrees, CKT is fundamental to instructional quality and significantly associated with students' learning gains. This study employs a novel analytic strategy that accounts for measurement errors in multilevel variance decomposition, providing clarity on the comparative sizes of within-school and between-school variation in teacher CKT. Results revealed that approximately half of the CKT variation lies within schools, while a third is across schools. Notably, although substantial within-school variation exists, it is largely random and not systematically associated with students' prior achievement or socioeconomic backgrounds. However, between-school variations indicate that schools with higher average student achievement and more advantaged socioeconomic backgrounds tend to have teachers with higher CKT.

The second study attempts to construct a comprehensive measure of instructional quality that integrates multiple instruments—observational ratings and student perception surveys—to analyze the causal relationship between CKT and instructional quality using a three-level

hierarchical linear model. This model accounts for the multilevel clustering of comparable classrooms within randomization blocks at the grade-by-school level, a unique experimental design feature of the MET project. Results found that that higher CKT significantly enhanced the mathematical quality of instruction, particularly in elementary and middle school classrooms. This is reflected in a significant increase in the richness of mathematical content. In contrast to mathematical quality of instruction, no significant CKT impacts were found on other dimensions of instructional quality, such as teacher-student interaction and student perceived classroom experience.

With enhanced measurements of CKT and multiple instruments of instructional quality, the third study examines the mediation pathways of instructional quality in the relationship between CKT and student achievement through a multi-step regression approach. This approach allows for a nuanced understanding of how different dimensions of instructional quality might mediate the effect of CKT on student outcomes. While high-school data are restricted by significant reductions in sample size due to missingness in administrative data, the mediation analyses identified noticeable indirect effects of CKT on student achievement through mathematical instructional quality in elementary schools and middle schools.

This dissertation provides robust evidence on the importance of CKT in shaping instructional practices and student achievement. By addressing critical gaps and employing rigorous analytical approaches, this research informs future educational policies and practices aimed at enhancing educational equity and effectiveness, ensuring all students receive high-quality instruction from knowledgeable and skilled teachers.

Keywords: Content knowledge for teaching, Mathematics teachers, Education equity, Instructional quality, Causal analysis, Causal mediation analysis, Student achievement

Chapter One. Overview

Ensuring every student has equal access to high-quality instruction have been longstanding goals of parents, educators, and policymakers. Many students from families of poverty, immigrants, racial minorities, and other marginalized backgrounds heavily rely on school instruction, as their families often cannot afford the supplementary learning resources outside of school (Alexander et al., 2004; Downey, 2023; Downey et al., 2004; Entwisle & Alexander, 1992, 1994). It is clear that teachers play a crucial role in the academic learning and future achievements for these disadvantaged students.

This study chose to focus on mathematics, considering that mathematics is a foundational subject essential for academic success across various disciplines in K-12 and postsecondary education. In the United States, mathematics education begins in kindergarten and continues through middle school, laying the groundwork for higher-level math and science courses in high school and college. Moreover, mathematics education often involves performance-based grouping and tracking, which can significantly impact students' academic trajectories, particularly for those from disadvantaged backgrounds (Figlio & Page, 2002; Gamoran & Mare, 1989; Mickelson & Everett, 2008; Vanfossen et al., 1987). Ensuring high-quality math instruction is therefore vital for promoting educational equity and preparing all students for future academic and career opportunities.

Shulman (1986) introduced the concept of teachers' content knowledge for teaching (CKT) as a comprehensive theoretical framework for teacher knowledge, integrating subject matter knowledge, pedagogical knowledge, and knowledge about curriculum, students and broader educational contexts. It is widely believed that teachers with high levels of CKT are better equipped to deliver high-quality instruction, thereby facilitating student learning and

improving academic performance. Furthermore, the effective transmission of teacher CKT heavily relies on the quality of teachers' instructional practice, suggesting that instructional quality may mediate the relationship between CKT and learning outcomes (Baumert et al., 2010; Kelcey et al., 2019; Kersting et al., 2012).

However, prior research on the relationships among teacher knowledge, instruction and student learning has been constrained by several factors. Firstly, teacher knowledge has conventionally been measured using qualification indicators such as certifications, advanced degrees, years of experience, under the assumption that teachers with these qualifications possess high CKT and are capable of deliver instruction of high quality. However, evidence showed some teacher qualification measures are at best weakly associated with student achievement and may not accurately reflect teacher knowledge required to deliver high-quality instruction (e.g., Goldhaber, 2008; Hanushek & Rivkin, 2006; Kane, Rockoff, & Staiger, 2008). Consequently, conclusions from prior research on systematic inequality of teacher knowledge and its impacts on student outcome are tentative. Furthermore, while theories and prior research have suggested that CKT plays a fundamental role in instructional quality and student learning (Baumert et al., 2010; Campbell et al., 2014a; Charalambous et al., 2020; H. C. Hill et al., 2005; Kelcey et al., 2019), comprehensive evidence, particularly utilizing a causal mediation framework, remains limited. This limitation is primarily due to the challenges of relatively small sample sizes in previous experimental studies and difficulties in accurately measuring variations in CKT.

Secondly, assessments of teacher knowledge and instructional quality are often inconsistent across studies, with measures designed or selected based on specific research focuses, restricting the generalization and integration of analytic results (Charalambous, 2020;

Mu et al., 2022). Additionally, many instructional quality measures fail to incorporate multifaceted constructs, potentially leading to findings with limited generalization.

Thirdly, prior research has largely focused on observational studies with minimal controls, primarily examining associations rather than establishing causality. Among a few studies that have explored causal links, many relied on professional development (PD) interventions that may not induce significant changes in CKT in the short term (Garet et al., 2011; Jacob et al., 2017; Roschelle et al., 2010; Santagata et al., 2010). Therefore, it is necessary to utilize the natural variation of CKT from a larger sample size to ensure sufficient heterogeneity for detecting the impacts of CKT.

I. Research Questions

This dissertation project aims to fill in the research gaps and generate rigorous evidence to answer three core research questions:

i. How does the distribution of teacher knowledge measured by CKT vary across classrooms and schools within a grade level? To what extent is this variation related to student composition?

ii. Do teachers with higher levels of CKT deliver higher quality instruction?

iii. Does the quality of instruction mediate the impact of CKT on student learning?

To answer these three research questions, the dissertation presents three empirical studies (Figure 1. 1) and is organized as follows: Chapter 2 examines how teacher CKT naturally varies across classrooms and schools within a grade level. Chapter 3 assesses the impacts of CKT on fine-grained dimensions of instructional quality using experimental data. Chapter 4 extends the line of mediation research by exploring the causal effects of CKT on student achievement via

various pathways of instructional quality. Chapter 5 briefly concludes and identifies areas for future investigations.

II. Research Design

This study utilizes the Measures of Effective Teaching (MET) longitudinal database, collected from a large-scale, two-year teacher evaluation project designed for a comprehensive investigation of effective teaching. The MET project selected six urban school districts across the United States, involving 2,741 teachers of English Language Arts (ELA), Mathematics, or Biology and their approximately 160,000 students in 4th to 9th grade classrooms. These school districts, along with the participating schools and teachers, were recruited through a process of "opportunity" sampling, meaning all participants volunteered. Consequently, the sample is not nationally representative, necessitating caution when interpreting and generalizing the results.

The MET project includes two survey waves: Academic Year 2009-2010 (year 1) and Academic Year 2010-2011 (year 2). The first year employed a purely observational design, collecting baseline data on districts, schools, teachers, and students, as well as videotaped classroom sessions rated by professional raters. Student outcomes were measured through state standardized test scores, supplementary tests, and student perceptions. More details can be found in Table 1. 1.

In the second year, the design incorporated an experimental component. School principals were asked to create "exchangeable" class portfolios at each subject-grade combination level, ensuring similar student composition across classes. The clusters of these "exchangeable" classrooms are called randomization blocks. If the experimental design is strictly implemented, any class-level differences in average prior achievement levels and class compositions are removed by design within the randomization blocks at the grade-by-school

level. Teachers eligible for the study were then randomly assigned to these exchangeable classes, forming the randomization blocks in year 2.

The sample size decreased from year 1 to year 2 due to school and teacher attrition (Table 1. 2). Specifically, 60 teachers from 11 schools were lost because their schools withdrew from the study. Additional teacher dropouts were due to transfers, leaves of absence, career changes, or changes in teaching assignments that made them ineligible for the second-year study. However, since randomization occurred after the first year, these dropouts should not affect the study's randomization process.

Focusing on Mathematics, the analytic sample for this study includes 913 Mathematics teachers and their 56,613 students in 267 schools for year 1, and 735 Mathematics teachers and their 12,209 students in 182 schools for year 2. This rich dataset allows for a comprehensive investigation into the distribution of teacher knowledge, instructional quality, and student learning outcomes.

III. Measurement

For measurement of CKT, this study employs well-developed assessment forms of teachers' subject matter knowledge and teaching skills from the six districts provided by the MET project and builds a measurement model built on item response theory (IRT). Unlike traditional measurement approaches, IRT models relax restrictions on the type of items used for measuring constructs and only assume that the latent ability follows a normal distribution in the population, an assumption that is highly likely valid for CKT assessments. The reliability estimate of the CKT latent scores is approximately 0.82.

To measure instructional quality, this study employed several commonly used instruments, including the Classroom Assessment Scoring System (CLASS), Mathematical

Quality of Instruction (MQI), and the Tripod student perception survey. Evidence from analyses aimed at uncovering the underlying structure across these instruments suggested against integrating information from various instruments when they measure distinct theoretical domains and perspectives of instruction. This is particularly relevant considering that the Tripod instrument is rated by students with year-long classroom experience, whereas other instruments are rated by trained professionals based on limited-time classroom observations.

IV. Analyses

Corresponding to the research questions, the analyses begin with a descriptive analysis of systematic inequality in the distribution of mathematics teachers' CKT. An uneven distribution of teacher CKT between schools and between classes within a school may contribute to educational inequality, which can be revealed by the multilevel variance decomposition of CKT. Preliminary findings show that teacher CKT, a direct measurement for teacher knowledge, is insignificantly and even negatively associated with conventional qualification indicators and value-added scores of teachers—a finding that diverges from common beliefs and prior research (H. C. Hill, 2010; H. C. Hill et al., 2011). Thus, analytic results of this study contribute novel evidence to the field.

The insights provided in the descriptive study naturally prompt an important theoretical question that will be investigated in the second study: How does teacher CKT influence instruction and subsequently student learning outcomes? Since primary teaching and learning activities occur during class, it can be presumed that teacher knowledge primarily affects student learning by influencing the quality of teaching in the classroom (Pianta & Hamre, 2009). Therefore, understanding the impacts of CKT on instructional practices is crucial for uncovering the mechanism by which various levels of CKT affect student learning.

Undoubtedly, the ultimate goal of improving teacher knowledge and the quality of instruction is to enhance student learning outcomes. With enhanced measurements of CKT and various instruments for instructional quality available, as well as large-scale experimental data from the MET database, the third study extends the investigation to identify how CKT affects student learning. This includes both direct pathways and through changes in specific aspects of instructional quality.

In summary, this dissertation seeks to address critical gaps in understanding the impact of teachers' CKT on instructional quality and student learning outcomes. By leveraging the rich, longitudinal data from the MET project, this research aims to provide robust evidence on three core research questions: the systematic inequality of CKT associated with student backgrounds, the causal impacts of CKT on the fine-grained aspects of instructional quality, and the mediating role of instructional quality in the relationship between CKT and student achievement. Through a combination of descriptive analysis and causal analyses, the subsequent chapters will present three studies, each corresponding to one of these research questions. Utilizing the improved CKT measurement, various instruments for instructional quality and extended causal analytic frameworks, this research not only contributes to theoretical advancements in the field but also offers practical insights for policymakers and educators striving to enhance the quality of education for all students.

Chapter Two. Inequitable Distribution of Teachers' Content Knowledge for Teaching Across Elementary and Secondary Math Classes

I. Introduction

Investigating the distribution of teacher content knowledge for teaching (CKT) is of great importance to educational equity. Students of low socioeconomic status (SES) heavily rely on mathematical instruction at schools, since they usually lack the learning support high SES students have at home that can supplement school education (Alexander et al., 2004; Downey, 2023; Downey et al., 2004; Entwisle & Alexander, 1992, 1994). Therefore, teachers play an even more important role in mathematical learning and future achievements for them than their peers from more advantaged backgrounds. However, evidence from prior research has found that teachers with better qualifications usually concentrated in schools with better funding. Such schools are generally located in higher SES neighborhoods and with a lower concentration of students from poverty, minority, and immigrant backgrounds. Consequently, if teacher qualifications are valid proxies for teacher knowledge and if the inequitable distribution of high-quality teachers continues, students from low SES backgrounds will be systematically deprived of equal opportunities to high quality mathematical education, which will ultimately sustain or even worsen the inequity in math achievement among students who differ in their family backgrounds.

Inequitable distribution of teacher knowledge may occur not just between but also within schools. Between-class ability grouping or tracking based on students' prior performance has been a common practice in K-12 mathematics education, especially in middle and high schools (Standing & Lewis, 2021). If schools purposely assign their most knowledgeable and effective teachers to classrooms consisting of high-performing students and leave low-performing students

with novice teachers, achievement gaps between low-performing students and high-performing students may continue to widen. On average, lower SES students tend to have struggled more in previous math learning and therefore display lower achievement. Therefore, under a tracking system, pairing high-quality teachers with high-performing students tend to further disadvantage low-SES students and further widen the SES-related achievement gaps. It would be nearly impossible for students who have performed relatively poorly in early grades to switch up to high ability groups, especially if these students have been constantly provided with a less advanced math curriculum and relatively low-quality instruction (Steenbergen-Hu et al., 2016). For this reason, inequitable teacher assignment may bear larger negative impacts if it happens in lower grades. Hence, it is important to investigate whether teachers' mathematical content knowledge is distributed inequitably at each grade level and whether inequitable distribution within a grade level is systematically associated with students' SES status.

Prior research has well documented disparity in exposure to qualified mathematics teachers between students from advantaged and disadvantaged backgrounds. However, two limitations are yet to be addressed: First, the majority of previous studies have used teacher qualification indicators, such as certification, teaching experience, college major coursework, and advanced degrees. However, some of these indicators, such as advanced degrees, are not math-specific and are weakly associated with students' learning outcomes, potentially limiting their construct validity as measures of teacher knowledge for math learning. Second, relatively few studies paid attention to disparities in teacher knowledge both within schools and between schools, where systematic sorting associated with students' prior academic performance and socioeconomic background may simultaneously occur (Chetty et al., 2013, 2014; C. T. Clotfelter et al., 2005; Conger, 2005; Goldhaber et al., 2015; Hanushek et al., 2005). With school-level and

classroom-level disparities in teacher knowledge under-studied, districts and schools would lack evidence to help improve the cultivation and allocation of teacher knowledge and further improve students' math learning.

Unlike prior research, this study utilizes latent scores of mathematical content knowledge for teaching (CKT), a measurement that directly assesses teachers' subject matter knowledge and teaching skills. Compared with conventional qualification indicators, teacher CKT is arguably a better alternative to estimate teacher effects, given that teacher knowledge is conceptually fundamental to instruction with high quality, and might have stronger impacts in student achievement. Importantly, this study reveals that teacher CKT is insignificantly, and even negatively, associated with conventional qualification indicators and value-added scores of teachers, a finding that diverges from common beliefs and prior research (H. C. Hill, 2010; H. C. Hill et al., 2011). Therefore, using teacher CKT, the evidence generated from this study contributes novel insights to the literature on unequal distribution of high-quality teachers.

Furthermore, this study employs a novel analytic strategy that accounts for measurement errors in multilevel variance decomposition and thus brings clarity to the comparative sizes of within-school variation and between-school variation of teacher knowledge. Specifically, the results reveal that after accounting for measurement errors that constitute around 16.7% of the total variance, half of the CKT variation lies within schools, and one third of it lies between schools. Grade levels, district fixed effects, and student composition mainly explain between-school variation. Although within-school variation is larger in proportion than between-school variation, there is no evidence of systematic sorting within schools. In contrast, systematic sorting between schools associated with students' disadvantaged status and prior achievement is apparent, particularly in high schools.

The organization of this chapter is as follows: The next section will lay out the theoretical framework. Section III will introduce the dataset and analytic strategy. Following that, Section IV will present the analytic results, and finally, Section V will conclude and discuss potential contributions.

II. Literature Review

1. Prior research on allocation of high-quality teachers

Variations in teachers' capability to deliver high-quality instruction naturally arise due to differences in their academic backgrounds, professional training, and teaching experience. However, if high-quality teacher is consistently unequally allocated and tends to favor those with pre-existing advantages, disadvantaged students may consistently find themselves taught by teachers with lower levels of knowledge and skills. Consequently, they may lag further behind their more advantaged peers in mathematics.

One strand of studies on disparities in access to high-quality teachers focuses on identifying teacher qualification gaps through investigating the disparities in exposure to qualified or highly qualified teachers between groups of students of different social origins, such as between historically marginalized minority populations and white populations or between economically disadvantaged students and their more advantaged peers (e.g., Cardichon et al., 2020; Corcoran, 2007; Knight, 2019). Through between-group comparisons, studies have revealed that systematic disparities exist in that students of low socioeconomic status (SES), with relatively low prior achievement, from historically underserved populations, and from immigrant families are in general taught by teachers that are considered less qualified (C. Clotfelter et al., 2004, 2008; Loeb, 2000; Springer et al., 2016; Steele et al., 2010). This phenomenon arises partly because schools serving higher proportions of disadvantaged students typically employ

teachers with less experience and lower qualifications because of budget constraints and relatively poor working conditions. These schools also face challenges in retaining qualified teachers who are competitive in the labor market, resulting in high turnover rates, which aggravates the disparities, Consequently, qualified teachers are concentrated in schools and districts that are well-funded, have good prior performance, and a high concentration of students from high SES backgrounds (Betts et al., 2000; Hanushek et al., 1999; Hanushek & Rivkin, 2012). This sorting of high-quality teachers between schools and districts leaves many students with disadvantaged backgrounds being taught by teachers with inadequate qualifications, resulting in persistent gaps in access to high-quality teachers.

In addition to systematic sorting between schools and districts, sorting of teachers into classes according to students' prior achievement, SES, and demographic backgrounds may also occur within schools (C. T. Clotfelter et al., 2006; Goldhaber et al., 2015; Kalogrides et al., 2011, 2013; Kalogrides & Loeb, 2013). Kalogrides and colleagues (2011, 2013) found evidence of within-school sorting where teachers of different qualifications and backgrounds were systematically matched with students' prior math and reading achievements from 4th to 11th grade levels (Kalogrides et al., 2011; Kalogrides & Loeb, 2013). Specifically, high-achieving students were more likely assigned experienced, white, male teachers, while low-achieving students were assigned less experienced, minority, female teachers. Teachers in leadership positions and those who had attended selective colleges were more likely assigned to teach high-achieving students. Moreover, there was evidence that the within-school sorting was also related to students' prior behavioral problems and attendance rates in record. Principals, parents, and even teachers themselves could be the reasons behind this matching pattern. For instance, principals may face conflicting priorities when assigning teachers: they may want to assign their best teachers to

students most in need while also needing to retain their most competent teachers. Additionally, concerned parents may pressure principals to match their children with teachers whose qualifications they find satisfactory. Moreover, teachers, particularly those who are more senior, may prefer to teach certain groups of students they find less challenging (Kalogrides et al., 2011, 2013). Goldhaber and colleagues (2015) examined the inequitable distribution of high-quality teachers utilizing teachers' licensure exam and value-added scores as well as teachers' experiences; identifiers of students' backgrounds include poverty status (eligible for free/reduced-price lunch), minority status, and prior academic performance. Evidence indicating inequities were universally found in all subgroup comparisons across elementary, middle, and high school classrooms in the State of Washington.

Additionally, Kalogrides and Loeb (2013) have found that sorting of qualified teachers may vary by grade levels, with teacher sorting being more prevalent in middle schools and high schools than in elementary schools. Specifically, between-school variation in teacher qualification and background is larger in elementary schools, while within-school variation tends to increase with grade levels (Kalogrides & Loeb, 2013).

Although sorting of qualified teachers occurs both between schools and within schools, there is no agreement about the comparative sizes of within-school sorting and between-school sorting (Chetty et al., 2013, 2014; C. T. Clotfelter et al., 2005; Conger, 2005; Goldhaber et al., 2015; Hanushek et al., 2005). Hanushek et al. (2005) maintained that the majority of the variation in access to high-quality teachers measured by value-added occurs within schools rather than between schools in 4^{th} to 8^{th} grades. Chetty et al. (2014) also found that over 85% of the variation in 4^{th} to 8^{th} grade teachers' value-added is within schools rather than between schools. Clotfelter et al. (2005) found that within-school or classroom effects explained

approximately one fourth of the total racial differences in exposure to novice teachers, slightly lower in proportions than district effects and school effects in 7th grade math classrooms. Goldhaber et al. (2015) examined teacher quality gaps in math and reading classrooms across 3th to 10th grades and concluded that most of the sorting of teacher quality measured by their qualifications came from district and school effects instead of classroom effects. Research contexts, methods, and various types of teacher quality indicators, as well as controls, may all affect the results. Nonetheless, regardless of where it occurs and their relative sizes, teacher sorting may have severe consequences for educational inequality, especially when students from the least advantaged backgrounds, e.g., racial minority groups, low SES, and immigration status, consistently taught by the lowest-quality teachers.

Following the line of scholarship that investigates systematic disparities in access to high-quality teachers associated with student backgrounds (e.g., Hanushek et al. 2005, Clotfelter et al. 2005, and Goldhaber et al. 2016), this study conducts variance decomposition of teacher CKT. However, instead of conducting a Blinder-Oaxaca decomposition analysis or constructing segregation indexes, this study uses hierarchical linear modeling that simultaneously analyze between-school variation and within-school variation, an approach that has been rarely employed in this literature. The results contribute new evidence on the multilevel variance of teacher knowledge measured by CKT.

Furthermore, this paper attempts to conquer the measurement limitation of previous research on teacher knowledge. Notably, prior studies used one or more qualification indicators as proxy for teacher knowledge, including teaching experience, certification, college rank, major or related coursework, and advanced degrees. However, given that some of these qualification indicators, e.g., advanced degrees weakly predict improvement in students' learning, the field

calls for evidence using alternative indicators of teacher knowledge that are strongly related to student learning. This study sets itself apart from prior research by employing a direct measurement of teacher knowledge--content knowledge for teaching. The advantages of using this direct measurement rather than qualification are discussed in the subsequent session.

2. Measurement of teacher knowledge in mathematics

Traditional definitions of "qualified" teachers, as outlined in federal education guidelines, often include criteria such as being certified, in-the-field, experienced, and/or holding a degree of master's or above. Hence, with little exception, prior studies have predominantly relied on one or a combination of these qualification proxies to assess teacher knowledge and its effects. However, an increasing body of recent studies has demonstrated that these qualification characteristics are only weakly associated, if at all, with students' academic performance (e.g., Goldhaber 2008; Hanushek & Rivkin 2006; Kane, Rockoff & Staiger 2008). In essence, as no conclusive evidence shows strong relationship between qualification and student achievements, the conventional understanding of "high-quality teachers", as indicated by teacher qualification does not necessarily align with the actual teacher effectiveness in influencing student learning outcomes (Hanushek & Rivkin, 2012). Given the limitations of these conventional qualification indicators, the evidence on inequity in access to high quality teachers remains tentative.

To overcome the limitations of qualification indicators, researchers have proposed using the outcome of teaching, i.e., students' academic achievement or learning growth as an alternative approach to estimate teacher effects. However, some of these outcome measures such as average standardized test scores in a class, year-to-year change in test scores etc., are imperfect to be used for estimating teacher effects. One challenge is to isolate the contribution of teachers to students' learning gains over time from other correlated contributing factors,

especially with the presence of nonrandom sorting of students into schools and classrooms (Baker et al., 2010; Hanushek & Rivkin, 2012; Kalogrides et al., 2013). Value-added measures represent a significant attempt to address this challenge by removing the influence of various observed confounding factors in assessing the effects attributable to teachers. However, researchers remain skeptical about relying solely on value-added measures in making important personnel or policy decisions. They argued that statistically, misspecification can occur because the value-added measure is calculated using a limited sample size of students and only a few observational time points, and cannot rule out influences from out-of-school learning and accumulating influences from previous years. Additionally, value-added measures may suffer from measurement errors, and lead to inference with flaws due to nonrandom sorting of students within and across schools. In practice, considering that value-added measures rely solely on standardized test scores, concerns also arise about teachers prioritizing teaching to the test rather than focusing on thoughtful and comprehensive instruction that meet student's diverse needs, potentially favor more advanced students and discouraging teacher collaboration (Baker et al., 2010; Corcoran, 2010). From these perspectives, value-added measures might not accurately indicate teacher effects.

This paper seeks to address the measurement limitations of prior research by focusing on the measure of content knowledge for teaching (CKT), also called pedagogical content knowledge, a concept incorporating both subject matter expertise and practical teaching skills. Unlike relying on qualification indicators or outcome-based measures that may not reliably measure teacher effects, CKT is arguably fundamental to instruction quality. A teacher's mastery of subject matter expertise and teaching skills substantially impacts students' knowledge acquisition (Shulman, 1986, 1987). Recent empirical evidence also suggests that CKT positively

predicts student learning gains (Baumert et al., 2010; Campbell et al., 2014a; Charalambous et al., 2020; Kelcey et al., 2019). However, it should be acknowledged that while CKT is a crucial component for high-quality instruction and student learning, it is not the sole determinant. Effective teaching also relies on other factors such as classroom management skills, the ability to engage and motivate students, and the availability of teaching resources. Therefore, while a strong foundation in CKT is necessary, it is not sufficient on its own to guarantee high-quality instruction and positive student outcomes. Additional elements and contextual factors may also play significant roles in the overall effectiveness of teaching and moderated the effects of CKT.

In this study, the assessment forms used to collect CKT data were collected using assessment forms adapted from the Mathematical Knowledge for Teaching measures developed by Ball, Hill, and colleagues (Ball et al., 2005, 2008; H. C. Hill et al., 2004, 2005; H. C. Hill, 2007; H. C. Hill et al., 2008; H. C. Hill, 2010), a measurement framework repeatedly validated in previous research. This paper constructs latent scores of CKT using item-level data from these assessment forms sourced from the Measures of Effective Teaching (MET) project database. To the best of my knowledge, no prior research has utilized latent scores of CKT constructed from an item-response-theory (IRT) model as a measurement of teacher knowledge for relevant analyses. By employing IRT-based CKT measurement, this study aims to provide more definitive evidence concerning the distribution of teacher knowledge in contrast with findings obtained through using qualification indicators or outcome-based measures.

3. Hypotheses based on prior research

Hypothesis 1. Between-school sorting of teacher CKT by student socioeconomic composition: Earlier research using qualification indicators such as certification, teaching experience, college major coursework, or advanced degree has found that schools with better

funding and located in higher SES neighborhoods attracted teachers with better qualification measured by these proxies. If some of these qualification indicators are positively correlated with teacher CKT, then schools with a higher level of socioeconomic composition of students may attract teachers with a higher level of CKT on average.

Hypothesis 2. Within-school sorting of teacher CKT by students' math preparation: Given the prevalence of between-class tracking and performance-based grouping in math based on pretest scores, which are positively associated with students' socioeconomic backgrounds, teachers of higher CKT are more likely to be assigned to teach classes in accelerated math tracks within a school.

III. Data

1. Sample description

This paper utilizes the Measures of Effective Teaching (MET) Longitudinal Database as the analytic sample. Supported by the Gates Foundation, the MET project collaborated with six urban school districts across the United States, gathering data on the teaching and learning activities involving approximately 2,700 teacher and their 160,000 students across 4^{th} to 9^{th}-grade English Language Arts, Mathematics and Biology classrooms. The database comprises comprehensive information from administrative records, including the demographic composition of classrooms, teachers, and schools, as well as students' standardized test scores from Academic Year 2004-2005 to Academic Year 2010-2011.

This paper primarily focuses on investigating the natural variation of CKT among Mathematics teachers, with relevant characteristics summarized in Table 2. 1. The analysis is based on data collected in Academic Year 2009-2010, a year before the implementation of an experimental design in the MET project.

2. Measurement

This study constructs latent scores of CKT as a measurement of teacher knowledge and examines its variation across classrooms, schools and districts. Latent scores of CKT are generated using models based on item response theory (IRT). Unlike traditional measurement approaches, IRT models relax restrictions on the type of items used for measuring constructs and only assume that the latent ability follows a normal distribution in the population, an assumption that is likely valid for CKT assessments. The reliability estimate of the CKT latent scores is approximately 0.82.

Notably, with common items across assessment forms, equating techniques can be employed to align the IRT latent scores obtained from each of the three assessment forms corresponding to different grade levels. However, it is important to recognize that a teacher proficient in teaching 9th-grade mathematics may not necessarily possess the pedagogical content knowledge required for teaching 4th-grade mathematics. Therefore, direct comparisons across grades are impractical. Subsequent analyses were based on subsamples segmented by school levels, with grade fixed effects included to ensure variations are analyzed within each grade level.

When inspecting the associations between CKT latent scores and teacher qualifications (see Table 2. 2), notable findings have emerged. Qualification proxies, particularly advanced degrees, exhibit a negative association with CKT scores, implying that conclusions regarding disparities in access to high-quality teacher may be inconsistent or even contradictory when relying on traditional qualification proxies instead of CKT measures. Furthermore, while a positive association between CKT and value-added scores is observed, the correlation is quantitatively small and statistically insignificant. These findings partially align with prior

research indicating that teachers' value-added and CKT levels can be mismatched. For example, teachers with high value-added may receive low ratings in the mathematical quality of instruction and low CKT. The new evidence also contradicts with previous evidence suggesting a positive correlation between teaching experience and CKT (H. C. Hill, 2010; H. C. Hill et al., 2011). The lack of strong linear correlations between CKT and conventional teacher qualification indicators suggests that the CKT measurement provides distinctive insights not found in previous studies. These results underscore the importance of examining disparities in access to high-quality teachers with the improved measure of CKT.

3. Analytic strategy

The analyses of this paper include decomposing the natural variation of CKT within the full analytic sample, as well as within elementary, middle and high school subsamples into the within-school component and the between-school component. It then examines whether teacher knowledge measured by CKT was distributed inequitably among students with different prior achievement levels or socioeconomic backgrounds. The analysis separates the dataset by school grade levels: elementary school grades (4^{th}-5^{th} grade), middle school grades (6^{th}-8^{th} grade), and high school grade (9^{th} grade). Parallel analyses are conducted within each of these three subsamples. Table 2. 3 presents comprehensive descriptive statistics of variables used in the analyses, categorized separately at the student, class, and school levels consistent with the original records in the MET database.

It's worth noting that the focus of this paper is on examining associations without establishing causation. Main variables of interest include teacher-level and school-level student composition, represented by percentages of students with specific disadvantaged statuses, including poverty (represented by eligibility for Free/Reduced lunch), minority (African

American or Hispanic), immigrant family (English Language Learner status) and disability (Special Education status), as well as students' average pretest scores in math (i.e., math score in 2009, one year prior to the MET project) at both the teacher and school levels. The values of the composition variables range from 0 to 1, with 1 indicating complete segregation. Pretest scores have been standardized and are considered comparable across different grades. During the Academic Year 2010-2011, some teacher participants taught multiple classes. Student composition at the teacher level is derived from all students taught by a given teacher. Additionally, when included in the regression analysis, teacher-level covariates are centered at their means within clusters identified by school ID, also known as group-mean centering. Through group-mean centering, coefficient estimates reflect changes in average teacher CKT levels associated with deviations in the profiles of the classes they teach from a typical class profile within their schools. Class profiles are defined by students' prior achievement levels and socioeconomic backgrounds. For example, with pre-test scores, group-mean centering ensures that the coefficient estimate represents the change in teacher CKT if the average pre-test score of a teacher's students is one unit higher than the average pre-test score of all students in that school. School-level covariates are centered at the overall sample mean, i.e., grand-mean centered. The coefficient estimates of these covariates represent changes in teacher CKT associated with a one-unit deviation of a school's student profile from the typical student profile across all schools.

The analyses primarily employ a three-level hierarchical generalized linear model (see model specifications in Table 2. 4). Given common clustering effects in school settings, utilizing a hierarchical linear model not only accounts for clustering that could otherwise compromise the validity of statistical inference, but also facilitates meaningful variance decomposition. A

hierarchical variation structure is appropriate especially when data are collected at multiple levels (Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002).

Notably, the availability of item-level data on teachers' CKT assessments in the MET database has provided an opportunity for researchers to address measurement errors, a challenge overlooked by prior studies. To overcome this methodological challenge, the regression models employed in this study has made several adjustments tailored to the current research context, which distinguishes itself from basic multilevel models. The primary adjustment involves incorporating a measurement model at the first level of the HGLM. This enables simultaneous consideration of unequal measurement errors across individuals while generating regression results. An alternative approach to address the unequal variance of measurement errors is by applying precision weights, calculated as the inverse of standard errors of empirical Bayes means for latent ability, estimated from a separate IRT model prior to the regression analyses. In essence, this alternative approach entails analyses based on a weighted two-level HLM at the second stage, with the first stage involving the estimation of CKT latent scores and measurement errors using the IRT model. This paper explores both approaches and will mainly discuss the results from HGLM in the subsequent sessions. The results of the weighted regressions are consistent with that of HGLM, reported in Appendix Appendix Tables

Table A2. 1 to Table A2. 7.

IV. Results

1. Variance decomposition of teacher CKT

The total variance of teacher CKT scores is 0.84. An analysis of variance decomposition (Table 2. 5) shows that approximately half (0.42) of the CKT variation lies within schools; around a third (0.28) is attributed to variation between schools; and the rest, roughly one sixth of

the CKT variation, is attributed to measurement errors. The estimated standard deviations of the measurement errors of CKT latent scores range from 0.32 to 0.63, with the average being 0.41.

When controlling for the grade and district fixed effects, the regression results of the unadjusted model across segmented subsamples by school levels shown in Table 2. 5 revealed that the ratio of within-school to between-school variation is approximately 10:1, consistent with findings from variance decomposition using value-added measures in several prior studies (Chetty et al., 2013; Hanushek et al., 2005). Including the grade fixed effects and district fixed effects, either separately or together, decreases the between-school variation. Furthermore, upon examining variance decomposition across models adjusting for different covariates, it becomes evident that including covariates mainly decrease the between-school variation (Table 2. 6). Essentially, the covariates and the fixed effects mainly explained between-school variation under the current research contexts.

This phenomenon regarding variance decomposition is consistently observed in the results from weighted regression. However, applying the precision weight (i.e., the inverse of the estimated posterior standard errors of the IRT latent scores, all below 1) will overestimate the magnitude of both within-school and between-school variations, particularly the former. Thus, this paper primarily relies on the results from the hierarchical generalized model. Future research should investigate the reasons behind the discrepancy in variance estimation between the two methods.

2. Systematic sorting of teacher CKT across schools

Analytic results revealed that the magnitude and significance of associations between teacher CKT and school-level students' prior achievement and socioeconomic statuses are more pronounced at higher grade levels (Table 2. 8 to Table 2. 12). This suggests that systematic

sorting of CKT between schools may become increasingly prevalent with higher grade levels. In high schools, all measures of student demographic composition exhibit significant associations with teacher CKT. Schools with a higher concentration of students from disadvantaged backgrounds tend to have teachers with lower CKT. Specifically, a ten-percentage-point increase of students with disadvantaged statuses taught by a teacher, such as poverty (represented by eligibility for Free/Reduced lunch), minority (African American or Hispanic), immigrant family (English Language Learner status) and disability (Special Education status) is associated with a decrease in teacher CKT by 0.166 (0.194 SD), 0.151 (0.176 SD), 0.278 (0.324 SD), 0.534 (0.623 SD) respectively, assuming linearity. The corresponding effect sizes (i.e., standardized regression coefficients, which will be used throughout the following paragraphs) are 0.044, 0.036, 0.054 and 0.027. This suggests a tendency for high schools to sort students based on their demographic backgrounds, a pattern that is noticeably distinct from that observed in middle and elementary schools.

Average math pretest scores at the school level significantly predict teacher CKT across elementary, middle, and high schools. Teachers with relatively higher CKT are concentrated in schools with students exhibiting relatively higher average pretest scores. The magnitude of the association increases with grade levels. Across elementary schools, a one unit (2.801 SD, detailed descriptive statistics can be found in Table 2. 3) increase in school average math pretest scores is associated with 0.248 (0.281 SD) increase in teacher CKT; across middle schools, the effect is 0.606 (0.654 SD); across high schools, it is 0.595 (0.694 SD). The corresponding effect sizes are 0.100, 0.268, and 0.227. One might suspect that teachers' relatively higher CKT in a school may have contributed to their students' relatively higher math pretest scores. However, 6^{th} graders and 9^{th} graders are new to middle school and high school, respectively. Their math

pretest scores could not have been affected by the CKT levels of the teachers in the middle schools or the high schools that they have just entered. Thus, there is clear evidence that teachers of relatively higher CKT are sorted into schools that enroll students of relatively higher prior math skills.

Systematic sorting of teacher CKT associated with school-level overall students' minority status (African American or Hispanic) is also significant across all three school levels. Specifically, a ten-percentage-point increase (0.360 SD) in proportion of students from minority background is associated with a decrease in teacher CKT by 0.035 (0.040 SD) in elementary schools, 0.088 (0.095 SD) middle schools, and 0.151 (0.176 SD) in high schools, assuming the association decreases proportionally. The corresponding effect sizes are 0.011, 0. 026 and 0.036. This evidence indicates that classrooms comprised with a higher concentration of students from minority backgrounds are taught by teachers with significantly lower CKT levels compared to their peers.

In summary, there is robust evidence on systematic sorting across schools, which is particularly pronounced in the 9th grade. Such sorting is significantly related to students' prior achievement and socioeconomic backgrounds.

3. Systematic sorting of teacher CKT within schools

The analytic results do not indicate systematic sorting within schools. Despite considerable CKT variation within schools, there is no evidence of systematical sorting based on students' prior achievement or socioeconomic backgrounds. Therefore, sorting within schools does not raise major concerns.

The only significant association observed is between the proportion of F/R lunch and teachers' CKT across classrooms within high schools. Unlike the association between school-

level proportion of F/R lunch and CKT, the association between teacher-level proportion of F/R lunch and CKT is positive, implying that in high-school classrooms with a high concentration of students eligible for Free or Reduced Lunch, teachers exhibit significantly higher CKT levels. Specifically, a ten-percentage increase (0.341 SD) in proportion of students eligible for F/R lunch taught by a teacher is associated with an increase in teacher CKT by 0.238 (0.278 SD), assuming the association decreases proportionally. The effect size is around 0.081. Results from weighted regression also supported this result. Despite its counterintuitive nature, it could be the case that this is a purposeful strategy of the high schools in these six urban districts to assign teachers to classrooms with a high concentration of children in poverty. Researchers may need to conduct qualitative research and gather additional evidence to explore the underlying reasons for this phenomenon.

Additionally, it should be noted that there is a significant negative association within schools between teachers' CKT and proportion of students eligible for F/R lunch across middle-school classrooms. However, this significant association is not found in the results obtained from weighted regression.

V. Conclusion and discussion

The paper contributes to the literature on disparities in access to high-quality teacher by utilizing latent scores of mathematical content knowledge for teaching (CKT) as a direct measure for teacher knowledge. Unlike conventional qualification indicators, such as advanced degrees, past research has suggested that CKT is fundamental to instructional quality and is significantly associated with students' learning gains. The study reveals that CKT is insignificantly, and even negatively, associated with conventional proxies including years of experience and advanced degrees, and with value-added scores of teachers, challenging common beliefs and prior

research. Analyses on unequal distribution of CKT add robust evidence and offer a fresh understanding of disparities in access to high-quality teacher, presenting new insights distinct from previous studies. Moreover, the study employs a novel analytic strategy that accounts for measurement errors in multilevel variance decomposition, providing clarity on the comparative sizes of within-school and between-school variation in teacher knowledge.

This study has revealed that substantial variation in math teachers' CKT lies within schools, even after controlling for the fixed effects of grades and districts, as well as covariates representing student composition. However, there is no evidence suggesting systematic sorting within schools associated with students' prior achievement and socioeconomic backgrounds. Essentially, within schools, students might be taught by teachers at different levels of CKT, but the variation in CKT is likely due to pure chance in classroom assignments and natural variation of CKT among teachers in current schools, factors unrelated to inequitable allocation of educational resources.

Another important finding is the systematic inequality between schools in student access to high CKT teachers. Notably, students' average prior math achievement emerges as a strong predictor of math teachers' CKT. Across all grade levels, teachers with higher CKT are concentrated in schools where students have higher average math pretest scores. This association increases in magnitude as grade levels rise from elementary school to high school. Moreover, from analytic results regarding the associations between teachers' CKT and school-level covariates, it is evident that the sorting of CKT between schools is systematically associated with students' socioeconomic backgrounds in addition to their prior math achievement levels. Such sorting is most pronounced in high schools, which is in alignment with evidence from previous studies (C. T. Clotfelter et al., 2002; Morgan & McPartland, 1981). In high schools, a high

concentration of students with disadvantaged statuses (i.e., status of poverty, minority, immigrant background, and disability) is significantly associated with low CKT levels of teachers. In essence, students in urban public high schools serving a large population of students from disadvantaged backgrounds tend to be taught by teachers with relatively low CKT levels. Systematic disparities in teacher CKT might affect the quality of mathematical instruction these high school students receive, consequently impacting their academic achievement and college attendance, which perpetuates severe education inequity.

This study adds to the large body of literature that studies systematic inequality in the distribution of high-quality teachers by employing a direct measure of teachers' mathematical content knowledge for teaching. The insights provided in the current study naturally prompt an important theoretical question: How does teacher CKT influence instruction, and subsequently, student learning outcomes? This question will be answered in the following two chapters.

Chapter Three. Does Teacher Content Knowledge Impact the Quality of Instructional Practices? A Causal Analysis

I. Introduction

Teacher knowledge is one of the few teacher characteristics that are significantly associated with students' learning outcomes (Baumert et al., 2010; Campbell et al., 2014a; Charalambous et al., 2020; H. C. Hill et al., 2005; Kelcey et al., 2019). In the study presented in the previous chapter, I explored the distribution of Mathematics teachers' content knowledge for teaching (CKT), a direct measurement for teacher knowledge specifically relevant to Mathematics. Since primary teaching and learning activities occur during class, it can be presumed that teacher knowledge primarily affects student learning by influencing the quality of teaching in the classroom (Pianta & Hamre, 2009). Therefore, understanding the CKT impacts on quality of instructional practices becomes crucial for uncovering the mechanism by which various levels of teacher CKT affect student learning. However, there is a lack of causal definition regarding the relationship between student exposure to a teacher with relatively high CKT and the instructional quality experienced by the student, let alone quantify this relationship. The unique experimental design of the Measures of Effective Teaching (MET) project provides an opportunity to investigate the causal relationship between teacher knowledge and instructional quality as rated by professionals and experienced by students.

Methodologically, the analytic strategy employed in this study sets a precedent for future studies analyzing multi-level data from experimental designs that randomize teachers to exchangeable classrooms within schools. The empirical results will have implications for improving the quality of instruction in underserved schools and classrooms by equitably allocating teacher resources. Notably, this study is not designed to answer the question about

whether interventions aiming at improving a teacher's CKT would lead to an improvement in the quality of instruction delivered by the teacher.

Furthermore, while some studies have explored how to incorporate the information across various classroom observation measurement frameworks (Blazar et al., 2017; Gill et al., n.d.; Lockwood et al., 2015; McClellan et al., 2013), very few have considered student perceptions, specifically in the context of mathematical instruction. This study extends this line of research by attempting to integrate professional observational ratings of classroom recordings with students' perceptions of their classroom experiences—a combination previously unexamined. By employing various instruments of instructional quality, including the Classroom Assessment Scoring System (CLASS), the Mathematical Quality of Instruction (MQI), and the student perception survey Tripod, this study provides a detailed view of the underlying structures within and across different aspects of instructional quality. The analytic results highlight the challenges in constructing a general instructional quality index through combining multiple instruments that measure distinct theoretical domains of instruction and represent different perspectives. Major distinctions include that the Tripod assessment used student ratings based on their year-long classroom experience, while the other instruments used ratings by trained professionals based on limited-time classroom observations.

With enhanced measurements of CKT and various instruments of instructional quality available, the study presented in this chapter intends to answer two research questions in the contexts of 4th- to 9th-grade mathematics classrooms across six urban public-school districts in US:

1) Do mathematics teachers with higher levels of CKT deliver instruction with higher quality on average than those with lower levels of CKT?

2) How are various instruments of instructional quality causally impacted by variations in mathematics teachers' CKT?

The subsequent sections of this chapter are organized as follows: Section II reviews the theoretical framework of various instructional quality measures and prior literature on the impacts of teacher knowledge on instructional quality. Section III presents hypotheses derived from the theoretical framework and prior knowledge. The next two sections provide a description of the data and present analytic results, including both factor analyses and causal analysis, followed by discussions and conclusions.

II. Literature Review

1. Measurement of instructional quality: using classroom observation and student perception

There is a lack of consensus regarding the definition and measurement of instructional quality in educational research. Some researchers focus on the product or outcome of instruction as the indicator for instructional quality. Average student achievement levels, or value-added scores in terms of student performance and other outcome-based measures are two most common indicators, where higher achievement levels indicate higher instructional quality (Chetty et al., 2013, 2014; Hanushek & Rivkin, 2010). Others emphasize the process of instruction, focusing on specific teacher behaviors and teacher-student interactions that align with theories of learning. Used in on-site classroom observation or video recordings analyses, observational protocols provide criteria for ratings and usually result in continuous scales, where higher ratings indicate higher instructional quality. (Brophy & Good, 1984; Cohen & Goldhaber, 2016). Additionally, some researchers argue that student perceptions are crucial components of instructional quality (Ferguson, 2012; Raudenbush & Jean, 2015; Scherer et al., 2016).

(1) Classroom observations

Classroom observations have been used as an assessment tool for teacher evaluation since the 1980s (Bell et al., 2019; Cohen & Goldhaber, 2016; Pianta & Hamre, 2009). Researchers believe that classroom observations have great potential for measuring teacher effectiveness and instructional quality, particularly in the context of advancing new common core standards, developing teacher evaluation systems, and promoting school policy changes (Cohen et al., 2022; H. Hill & Grossman, 2013; Liu & Cohen, 2021). Observational ratings, rich in detailed information and relatively easy for teachers to interpret, may provide actionable feedback and pinpoint specific areas for instructional improvement. The feedback can help teachers make informed adjustment to meet the increasing accountability requirements (Bell et al., 2019; Cohen et al., 2022; Cohen & Goldhaber, 2016; Pianta & Hamre, 2009).

Researchers have designed numerous observational protocols for teacher evaluation, such as the Classroom Assessment Scoring System (CLASS), the Framework for Teaching (FFT), the Mathematical Quality of Instruction (MQI), and the Protocol for Language Arts Teaching Observations (PLATO). While all these protocols measure instructional quality, they differ in theoretical foundations, measurement constructs, subject focus, implementation, and other aspects. To illustrate, I mainly review two protocols. The first is CLASS, an observational protocol emphasizes fostering a supportive environment and providing affirmation to students that can be generally applied to classrooms of all grades and subjects (Pianta & Hamre, 2009); the second is MQI, an classroom observational instrument targeting mathematics and specifically designed for mathematics instruction (Learning Mathematics for Teaching Project, 2011).

CLASS. CLASS is an evaluation system based on standardized and validated observational protocols for assessing teacher-student interactions occurring in classroom.

Aligned with developmental theories, CLASS includes three major domains for assessing classroom instructional quality: emotional support, classroom organization, and instructional support. Each domain comprises specific sub-domains.

Pianta and colleagues argue that an effective teacher's role in classroom extends beyond the teaching of content to include socializing, motivating, and mentoring students, that teacher-student interactions, particularly those that nurture a positive learning environment and manage behaviors effectively, significantly and positively contribute to students' social and academic development. These observable teacher behaviors are central to quality instruction and teacher effectiveness. A standardized and validated assessment of classroom interactions, if tested and proved as consistent, should be included in accountability frameworks, teacher preparation and professional development, national surveys, as well as value-added measures (Hamre et al., 2007, 2013; Hamre & Pianta, 2007; Pianta et al., 2011; Pianta & Hamre, 2009).

Another commonly used observational instruction protocol FFT also centers on teacher-student interactions, emphasizing their role as determinants of teacher effectiveness (Cohen et al., 2022; H. Hill & Grossman, 2013; Liu & Cohen, 2021). While CLASS is based on developmental learning theories, FFT is grounded in constructivist learning theory and emphasizes teachers' professional responsibilities both within the classroom setting and outside the classroom (Danielson, 2007, 2008; Danielson & Axtell, 2009). Content-general instruments such as CLASS and FFT assume that teachers' behaviors that facilitate or enhance teacher-student interaction are universally applicable across all grades, subjects, and educational contexts. However, this assumption may not always hold true.

MQI. MQI is a protocol specifically designed for evaluating Mathematics instruction. Its subject-specific theoretical framework distinguishes it from subject-general instruments focusing

on pedagogy. According to the MQI developers, "quality" is represented by distinctive features of instruction; "instruction" is defined as dynamic interactions involving teachers, students, and content; and "mathematical" refers to particular knowledge in the discipline of mathematics relevant to teaching (Learning Mathematics for Teaching Project, 2011). Through analyzing three sources of evidence—prior literature, experience of teaching and studying teacher education, and classroom videotapes of mathematics classrooms over an entire academic year—developers of MQI established its framework, encompassing seven major theoretical constructs with point rating scales. They also identified missing elements of MQI that could not be turned into reliable constructs, such as the launch of mathematical tasks and the scaffolding of student work (H. C. Hill et al., 2008, 2012, 2018; Learning Mathematics for Teaching Project, 2011). In a multiple-case analysis that aims to systematically discuss strengths and limitations of using MQI to measure instructional quality, Charalambous and Litke pointed out that while providing rich information of discipline-relevant instructional practice, MQI as a content-focused instrument is inevitably limited in capturing generic pedagogy aspects. They suggested that utilizing MQI complementarily with instruments that are content-general may capture and explain more variation in teachers' instructional quality than by each instrument separately, an argument consistent with other studies' conclusions (Blazar et al., 2017; Charalambous & Litke, 2018).

(2) Student perception

As actual participants in classroom instruction, students provide ratings based on their year-long experiences, offering valuable insights that external observers, who only see a brief segment of instruction, might miss. Some researchers even argue that student perception is one of the most important criteria for assessing teacher quality and effectiveness (Ferguson, 2012; Polikoff, 2015; Scherer et al., 2016; Wagner et al., 2013). Various formats exist for capturing

student perceptions, ranging from large-scale surveys to focus group interviews. Among these, student questionnaires about their classroom perceptions are widely used because it is easy to implement and is cost-effective in terms of resources and time. However, ratings based on student self-report are often criticized for their instability and lack of reliability, as the ratings can be biased if students are influenced by subjective factors such as a teacher's popularity, the difficulty of the subject, or the student's own interests, rather than focusing on the quality of the actual instruction (Gitomer, 2019; Senden et al., 2022; Wagner et al., 2013).

Here, I primarily review research based on the Tripod survey, a widely used student-centered survey that assesses instructional quality based on student perceptions. A research team led by Ferguson defined seven theoretical dimensions of the Tripod survey as the 7Cs: Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate (Ferguson, 2012; Phillips et al., 2021; Rowley et al., 2019). Using the 7Cs framework, Raudenbush and Jean (2015) identified Control and Challenge as two critical aspects of instructional quality that result in significant learning gains (Raudenbush & Jean, 2015). Subsequent research found that the 7Cs framework does not always best fit the data, likely due to varying research contexts. For instance, Wallace et al. (2016) found that a bi-factor structure, consisting of a general factor and classroom management, provided a better fit to their data. Similarly, Kuhfeld (2017) identified a two-dimensional structure, comprising a Control factor and an Academic Support factor that includes the other 6Cs, as sufficient to explain her data on 6th-8th grade math and ELA teachers in secondary schools (Kuhfeld, 2017; Wallace et al., 2016).

Despite criticisms regarding subjectivity and instability, student perception is considered to have incomparable advantages and has been widely accepted a tool for assessing instructional quality. A large body of literature showcases researchers' attempts to address the validity

concerns and measurement challenges (Kuhfeld, 2017; Wagner et al., 2013; Wallace et al., 2016). With both student perception ratings and observational ratings from external raters available in the MET database, this study aims to explore the possibility of increasing validity by systematically integrating objective and subjective ratings.

(3) Cross-instrument structure of instructional quality

A review of the underlying theories and domains across various assessment instruments for instructional quality reveals that different approaches capture distinct facets of instructional quality, each with its advantages and limitations. These instruments can measure both common and complementary constructs while maintaining their distinctive elements (Berlin & Cohen, 2018; Praetorius & Charalambous, 2018). Exploring how to systematically combine these constructs across instruments to generate a comprehensive measurement of instructional quality is both theoretically and methodologically important and holds great potential (Schlesinger & Jentsch, 2016). However, districts and schools often face constraints in funding and human resources, making it impractical to use multiple observational protocols to evaluate teachers' instructional quality. Consequently, most districts employ only one protocol. This limitation has prevented researchers from accessing data containing ratings from multiple measurement tools for instructional quality, as such datasets did not exist.

Thanks to extensive collaborations among researchers across the US and support from the Gates Foundation, the MET database has made ratings from multiple measurement protocols available for the first time. More importantly, these ratings are based on video recordings submitted by the teachers and questionnaires completed by the student cohorts they taught during the 2011-2012 academic year. This rich dataset provides a unique opportunity for researchers to study the underlying structure across existing observational protocols, determine if they measure

common and complementary constructs, and generate a comprehensive picture of classroom instructional quality by incorporating information from various ratings. For instance, Gill et al. (2016) conducted a content analysis (qualitative coding) through reviewing the detailed rubrics of five commonly used instruments available in the MET database: the CLASS, the FFT, the MQI, the PLATO, and the UTeach Observational Protocol. They identified potential overlaps among the assessment items from these various instruments and classified them into cross-instrument dimensions. Their analysis revealed that eight out of the ten dimensions of instructional practice are common across all five instruments, indicating a high level of conceptual consistency (Gill et al., 2016).

Lockwood et al. (2015) employed a novel Bayesian exploratory factor analysis method to investigate the structure of effective teaching using classroom observational ratings from the MET database for Mathematics and English Language Arts (ELA) separately. Their analysis of items from MQI, CLASS, and FFT revealed two primary cross-instrument factors, "Instructional Practices" and "Classroom Management," which together comprise the main construct of effective teaching. The cross-instrument structure with two primary factors is consistent for Math and ELA (Lockwood et al., 2015).

Blazar and colleagues (2017) conducted extensive exploratory (EFA) and confirmatory factor analyses (CFA) to investigate cross-instrument factors between the CLASS and MQI observational ratings. Their exploratory factor analysis revealed clusters of items that span multiple instruments, although the degree of overlap was less pronounced compared to the findings of Lockwood et al. (2015). They termed these cross-instrument factors "instructional factors" to distinguish them from "instrument factors" composed of items within the same

observational protocol. Furthermore, they explored various CFA models, including those with and without a bi-factor structure (Blazar et al., 2017; Lockwood et al., 2015).

The optimal CFA models identified by Blazar et al. (2017) included "instructional factors," indicating the presence of cross-instrument factors, alongside two "instrument factors," each respectively corresponding to subject-specific protocol and subject-general protocol. Based on these findings, they suggested that studying teaching, a phenomenon characterized by a multidimensional and conceptually complex structure, necessitates a comprehensive construct that integrates both subject-specific and subject-general instruments (Blazar et al., 2017).

Prior research efforts to construct cross-instrument factors have revealed overlaps among various instruments of instructional quality (Blazar, 2015; Blazar et al., 2017; Gill et al., 2016; T. Kane, 2012; T. J. Kane & Staiger, 2012a; Lockwood et al., 2015). Additionally, even when using the same sets of observational scores from the same source, different contexts can influence the best-fit construct for measuring instructional quality, often diverging from the original constructs proposed by instrument developers. These findings highlight the necessity of conducting factor analysis to examine the underlying structure. Such analysis may uncover a parsimonious structure that can enhance subsequent analyses both conceptually and empirically.

Notably, no prior research has examined the underlying structure of both classroom observational ratings by professional raters and student-perceived classroom experience surveys. This study aims to fill this gap by exploring the practicability of a multi-dimensional construct of instructional quality, incorporating both observational ratings and student perceptions. By doing so, it seeks to capture instructional quality more comprehensively than previous studies. The resulting measurement framework aims to integrate common and complementary constructs across instruments while preserving their distinctive elements.

2. Relationships between CKT and instructional quality

Although researchers generally anticipate that high levels of CKT would lead to high quality of instruction, few studies have empirically investigated the direct links from teacher content knowledge to ratings on instructional quality using various instruments.

The exception is the MQI. As both MQI and the instrument for assessing Mathematical (Content) Knowledge for Teaching were developed by the same research team, they had the tools and data to gather evidence on how mathematical CKT is associated with the quality of instruction within the context of mathematics education, examining fine-grained constructs specifically related to teaching Mathematics. Hill et al., (2008) found a positive and strong correlation between mathematical CKT and MQI scores. Their exploratory case study highlighted that teachers with higher CKT levels were better at avoiding mathematical errors, appropriately responding to students' mathematical expressions, and addressing misunderstandings. Additionally, they identified two contextual factors that could mediate this relationship: teachers' beliefs about making math classes enjoyable for students and their adherence to curricular materials. Subsequent analyses confirmed these significant associations between mathematical CKT and MQI (Charalambous, 2010; H. C. Hill et al., 2008, 2011). However, these studies cannot rule out student composition as a competing explanation for the association. It is possible that higher-achieving students are more likely to be assigned to teachers with higher mathematical CKT and that instruction may appear to have a higher quality when students are better prepared.

Prior studies on teacher-student interactions in the classroom have predominantly focused on how the quality of these interactions affect student outcomes rather than examining how various levels of teacher CKT might influence these interactions; and even less research has used

high-quality instruments to examine this relationship. This gap is also present in studies on student perception and the Tripod survey. As previously mentioned, researchers did not have access to instructional quality ratings from multiple observational protocols before the MET project. Consequently, no prior research has directly linked CKT with CLASS or Tripod scores.

Beyond specific instruments and towards a broader definition of instructional quality, the field lacks substantial evidence on the relationship between CKT and instructional quality. Some relevant findings can be found from studies examining relations of CKT, instruction and student outcomes, albeit their primary emphases were on how CKT and instructional quality affect student outcomes (Baumert et al., 2010; Copur-Gencturk, 2015; Kelcey et al., 2019; Kersting et al., 2012).

Baumert et al., (2010) found significant effects of CKT on instructional quality measures regarding "Curricular level of tasks", "Cognitive level of tasks", and "Individual learning support" while no significant effect of CKT was observed for "Classroom management". Kersting et al., (2012) used their own CKT and Instructional quality measures and reported that one standard deviation (SD) increase in CKT measure was associated with a two-thirds SD in overall instructional quality score. Kelcey et al., (2019) found that one SD change of CKT is associated with 0.22 SD change in Ambitious Mathematics and Errors domains derived from MQI, but not with CLASS domains, Ambitious General and Classroom Organization. Copur-Gencturk (2015) observed that improvement in teachers' mathematical CKT through intensive training on inquiry-based teaching correlated significantly with changes towards a more meaning-making teaching agenda, and a positive classroom climate, while no statistical relationship was found with the mathematical quality of classroom tasks or student engagement. In contrast, Shechtman et al., (2010) did not find statistically significant relations between

teachers' CKT levels and their decision-making regarding classroom instruction, including the complexity of the topics focused in class, the complexity of their teaching goals and the use of technology such as time allocated in computer labs. In summary, the evidence is mixed, and the inconsistent measurements of CKT and instructional quality hinder comparison across studies (Baumert et al., 2010; Copur-Gencturk, 2015; Kelcey et al., 2019; Kersting et al., 2012; Shechtman et al., 2010a).

The lack of solid evidence on the relationship between CKT and instructional quality necessitates a thorough analysis, particularly concerning the relations between CKT and both observational teacher-student interactions and classroom experiences as perceived by students. The MET database is unique as it employed multiple instruments to provide multi-dimensional measurement of instructional quality; moreover, by randomly assigning teachers to exchangeable classrooms within each school, the experimental design rules out the confounding impacts of student composition. Thus, this study will contribute valuable insights to the field. The results will be useful for future research aiming to establish a more comprehensive evaluation framework for teacher performance, providing interpretable and actionable feedback, and improving the design of future professional development programs in practice.

III. Hypotheses

Hypothesis 1. MQI: When teaching comparable classes of students, mathematical teachers with a higher level of CKT are expected to have higher ratings on MQI on average. Evidence from prior research has suggested that the impacts might be prominent in certain sub-dimensions, such as reducing the frequency of errors and imprecision as well as being able to help student participate in meaning making and reasoning.

Hypothesis 2. CLASS: When teaching comparable classes of students, mathematical teachers with a higher level of CKT are expected to have higher ratings on CLASS on average. This is because a higher level of content knowledge for teaching allows teachers to better maintain a positive classroom environment, including showing more respect for students, being more patient with mistakes and classroom management, and allocating more time to provide additional feedback.

Hypothesis 3. Student perception: When teaching comparable classes of students, mathematical teachers with a higher level of CKT are expected to generate more positive perceptions of classroom experiences among their students. Their expertise in the subject matter can help them earn students' respect, increase students' interest in mathematics, and maintain a welcoming and supportive classroom environment. Additionally, high CKT often correlates with effective pedagogical strategies, making students feel more comfortable and engaged in class. These factors collectively contribute to more positive student ratings.

Hypothesis 4. Variation by school level: The impact of CKT on instructional quality can vary across school levels due to differences in students' developmental stages and educational contexts. Specifically, in elementary and middle school classrooms where the math curriculum is less demanding, teachers with high CKT may prioritize interactional behaviors. These behaviors help scaffold learning and create a supportive environment to foster student interests and facilitate confidence building. Consequently, the impact of teacher CKT on classroom interaction and certain dimensions of MQI are likely evident in these settings. In high school classrooms, the curriculum demands a higher level of mathematics-focused instruction rather than extensive teacher-student interaction. Additionally, adolescents with growing self-awareness and self-concept are likely more perceptive of the impact of teacher CKT on their learning experience

when compared to younger children in lower grades. This is likely due to adolescents' heightened ability to reflect on and internalize the quality of instruction they receive. Therefore, the influence of CKT on instructional quality might vary significantly by grade level due to different developmental stages of students and changing emphases by the curricular requirements.

IV. Data description

Overview. The analytic data were derived from the Measures for Effective Teaching (MET) study, a large-scale two-year teacher evaluation project designed to investigate, identify, and comprehensively measure effective teaching skills and practice. The MET project recruited six urban school districts across the United States, with a total of 2,741 teacher volunteers participating. Among these teachers, this paper focuses on the 735 teachers teaching Mathematics during the Academic Year 2010-2011 when the MET project conducted randomized experiments. The analytic dataset was constructed using main files at the class session level, merged by unique IDs for teachers, schools and districts; the data set includes item-level observational scores for three classroom observation protocols used to construct instructional quality measures: MQI, CLASS, and Tripod.

To distinguish between different developmental stages and educational contexts, I conduct parallel analyses for three subsamples categorized by grade levels separately: elementary schools, middle schools, and high schools. In the elementary school sample, there were 324 teachers each assigned to a classroom within 121 randomization blocks across 80 schools. The classrooms within the randomized blocks are structured to ensure exchangeability in terms of classroom compositions including students' academic and socioeconomic backgrounds. The middle school sample consisted of 257 teachers in 76 schools, each assigned

to a classroom in 117 randomization blocks. The high school sample comprised 83 teachers in 35 schools and 38 randomization blocks. I have excluded the randomization blocks with only one class section.

Student perception survey. The Tripod survey used in the MET project came from the survey forms designed by Ronald Ferguson of Harvard University. Ratings of students' perceptions were collected and measured using five-point Likert scales (ranging from "Totally Untrue" to "Totally True"). The Tripod questionnaire comprised a total of 49 items, covering students' perceptions of their classroom experience, as well as survey questions regarding their demographics and socioeconomic backgrounds at home. Elementary school students were administered a shorter form of the survey compared to that for secondary school students. Based on previous research, I identified 19 items relevant to the domains of 7Cs in the elementary school data and 34 items in the secondary school data, as detailed in the Table A3. Within each sub-sample, I average Tripod's itemized rating scores to obtain class-level mean scores using unique class section IDs.

Classroom observation measures. The MET data employed professional raters to evaluate classroom video recordings using five observational protocols widely accepted in practice. Among these, three were subject-specific: the Protocol for Language Arts Teaching Observations for ELA; the Quality Science Teaching Sciences Instrument for Science; and the Mathematical Quality of Instruction (MQI). The rest of the two protocols, the Framework for Teaching (FFT) and the Classroom Assessment Scoring System (CLASS), were designed to be universally applied across all subjects.

After inspecting the codebooks of the observation scores from the MET database, I notice that a general instrument and a subject-specific instrument may be complementary to each other

and may have some degree of overlapping across their measurement items, consistent with the assumptions in prior studies. FFT and CLASS both measure the general classroom atmosphere. However, FFT additionally includes domains and items that assess instructional practices occurring outside the classroom, such as planning and preparation before class, and designing and analyzing assessments after class. To ensure the focus of the instructional quality measure is on practices occurring during the class period, and to align the subject-general instrument with the other two sources of ratings (subject-specific instrument for in-class instruction and student perception surveys of classroom experience), this study chooses MQI instead of FFT. Future extensions of the study may include FFT along with other measures to explore the validity of creating a broader measurement of instructional quality.

To recap, this study to attempts to construct a multi-dimensional measurement for instructional quality that could incorporate information from these three sources of classroom ratings, i.e., the subject-specific measure MQI, the subject-general measure CLASS, and the student perception measure Tripod. The domains originally defined in the three rating frameworks can be found in Appendix (Table A3. 1 to Table A3. 3).

V. Analytic results

1. Factor analysis on classroom ratings

The factor analysis in this paper adopts the strategy outlined by Blazar et al. (2017) who also analyzed the MET data for cross-instrument factors. However, this analysis differs from theirs in several key respects. Notably, while their study focused on examining the CLASS and the MQI observational instruments, this research also encompasses Tripod survey, which rated the classroom from the perspective of students. Additionally, whereas Blazar et al. concentrated on 4th to 5th grade teachers, this study encompasses a broader range of teachers in classrooms

spanning from 4th to 9th grades. It is important to note that while the CLASS protocol is uniform across all grade levels, the MQI and Tripod surveys were administered separately to accommodate the distinctions between elementary and secondary education contexts. To maintain consistency with the original design of the MQI and Tripod protocols, I segmented the analytical dataset into two sub-samples: one comprising elementary school grades (4^{th}-5^{th} grade) data and another encompassing middle and high school grades (6^{th}-9^{th} grade) data. Furthermore, unlike Blazar et al., who had access to detailed item-level scores of the MQI, the MQI data available to this research are less fine-grained.

(1) Exploratory factor analysis

The objective of the exploratory factor analysis is to unveil the underlying patterns of intercorrelation among classroom observation measures within the current dataset, providing evidence to guide decisions regarding the optimal number of factors needed to explain the instructional quality measures. To achieve this goal, I inspected the pairwise correlations of all items within and across observation protocols, along with their initial or unrotated factor loadings.

a. Analysis within each framework.

Classroom Assessment Scoring System (CLASS). Pairwise correlations among items within CLASS are moderate, ranging from 0.21 to 0.76 in the elementary school sample and relatively stronger in the middle and high school sample, ranging from 0.31 to 0.80. Among the CLASS items, Negative Climate was the only item negatively associated with others.

The patterns of factor loadings for CLASS itemized scores in the current dataset did not conform to the four higher-level conceptual dimensions of CLASS. In middle and high school data, items primarily loaded into two factors, which explain cumulatively 98.6% of the test

variance. All items load onto factor 1 (Eigenvalue=7.74, Difference=6.44, proportion=0.844). Negative Climate, Behavioral Management and Productivity loaded onto factor 2 (Eigenvalue=1.30, Difference=1.04, proportion=0.142) with a substantial loading (> 0.4 in absolute value), which is consistent with Blazar et al. (2017). The pattern in the elementary school data closely resembled that of the secondary school sample. Detailed results can be provided upon request.

Mathematical Quality of Instruction (MQI). Correlations among MQI items are shown in Table 3. 1. Initial loadings of MQI items in middle and high school data result in only one factor with an eigenvalue greater than 1 (Eigenvalue=1.26, Difference=0.85, proportion=1.04). Only three items have a substantial loading (>0.4) on this factor, which are Student Participation in Meaning Making and Reasoning (SPMMR), Richness of Mathematics, and Working with Students and Mathematics (WWSM). Detailed results can be found in Table 3. 2. The rest of the three items have small loading, e.g., -0.09 for Error and Imprecision, which are considered not substantial. The pattern in the elementary school data is consistent with that of secondary school sample.

Tripod. Given that the Tripod survey does not include items regarding negative perception in its elementary school form, the correlations are all positive and range from 0.08 to 0.63. In the middle and high school sample, items relevant to negative perception of the classroom experience or teachers' behavioral management practice are included. The correlations among items in Tripod's secondary school form are stronger, ranging from 0.31 to 0.80 and with negative correlations ranging from -0.47 to -0.0026.

Initial factor loadings of Tripod in the current dataset suggested no clear item-clustering pattern akin to the original 7Cs domains. Instead, the number of factors required to explain

students' perception of classroom experience is relatively small. In the elementary school data, all 19 items loaded onto the first factor, which is the only factor that has an eigenvalue greater than 1 (Eigenvalue=6.45, Difference=5.75, proportion=0.818). In the middle and high school data, Factor 1 (Eigenvalue=16.34, Difference=13.04, proportion=0.717) was measured by items relevant to positive perceptions, Factor 2 (Eigenvalue=3.31, Difference=1.69, proportion=0.145) by items for negative perceptions, and Factor 3 (Eigenvalue=1.62, Difference=0.81, proportion=0.071) by items relevant to teachers' classroom management that might lead to negative perceptions, some of which already exhibited substantial loadings on Factors 1 and 2. Detailed results can be provided upon request.

Across each instrument, the initial factor loading results were consistent with the Cronbach's alpha of each instrument. Scale reliability coefficients of CLASS and Tripod exceed 0.90, while those of MQI remain lower than 0.50, for both elementary and secondary school data. The Cronbach's alpha results indicated excellent internal consistency for CLASS and Tripod, while MQI demonstrated a low level of internal consistency. Consequently, it might be reasonable to construct CLASS and Tripod measures as a whole unit, while MQI items need to be analyzed separately.

b. Analysis across observational frameworks.

The correlations between items from different observational frameworks tend to be weak in general, indicating minimal overlapping between these measurement frameworks. Specifically, pairwise correlations between MQI and CLASS items typically range from -0.16 to 0.39, with most falling below 0.2. The correlations between MQI and Tripod items are even weaker, ranging from -0.11 to 0.13 in elementary sample, and from -0.19 to 0.29 in secondary sample. Meanwhile, correlations between Tripod and CLASS items range from 0.21 to 0.24 in

elementary sample and from -0.34 to 0.43 in secondary sample, representing at best moderate correlations. This suggests that they might capture different aspects of instructional quality, which is to some extent inconsistent with previous research.

The patterns of factor loadings for CLASS and MQI items remain consistent with the results obtained from analyzing each instrument separately. These items primarily load onto and cumulatively measure specific factors, even when items from other instruments are included in the analyses. For example, upon examining the initial loadings of all items from CLASS, MQI, and Tripod in secondary school data, the third factor predominantly comprises negative items from Tripod, aligned with the second factor observed in the initial loadings of Tripod data alone.

However, there is evidence suggesting potential existence of overlapping or cross-instrument factors. For instance, upon examining the initial loadings of CLASS and MQI items in secondary school data, the first factor is measured by all CLASS items and three MQI items (SPMMR, Richness, and WWSM), all of which exhibit substantial loadings. Similarly, the first factor (Eigenvalue=21.82, Difference=16.02, proportion=0.439) identified by the initial loadings of all items from CLASS, MQI and Tripod in secondary school data consists of all items from CLASS and all positive items from Tripod. Furthermore, the second factor (Eigenvalue=5.80, Difference=1.46, proportion=0.117) is measured by all CLASS items and three MQI items, SPMMR, Richness, and WWSM, the clustering of which is also found in the results examining factor loadings patterns across MQI and CLASS items.

In summary, the structures unveiled by exploratory factor analyses within each instrument, especially for CLASS and Tripod, should be prioritized as the primary structure when constructing the final Structural Equation Models (SEMs). Meanwhile, the clustering

revealed by factor loadings across instruments suggest that cross-instrument factors should also be examined in confirmatory factor analysis.

(2) Confirmatory factor analysis

To ensure the robustness of dimension construction, I fit different Structural Equation Models (SEMs) based on evidence derived from the original conceptual framework, prior studies (e.g., Blazar et al., 2017 and Gill et al., 2016), and exploratory factor analysis. The overarching strategy involved initially testing the original conceptual dimensions due to their strong theoretical underpinnings, assessing whether these dimensions formed a suitable structure for SEM in the current dataset. Subsequently, adjustments were made to the structure based on insights from previous research, and further refinements were considered using evidence from exploratory factor analysis to enhance fit, particularly when convergence was not achieved. It is noteworthy that the MQI does not have a higher-level structure of domains beyond the available ones.

The confirmatory factor analysis has provided evidence that, in the current dataset, using composite scores of the CLASS and Tripod to construct instructional quality measures is acceptable. For instance, in the elementary data, I conduct confirmatory factor analysis of the CLASS itemized scores with increasing structural complexity, including structures without higher-level factors, structures with the four theoretical dimensions each constructing a second-level factor, and a structure incorporating four second-level theoretical dimensions and a third-level factor measuring overall instructional quality. Although these SEM-constructed scores exhibited variations in means and variances, their pairwise correlations exceeded 0.99, indicating a perfect correlation. Additionally, the correlations between SEM-constructed domain factors were strongly correlated (> 0.80 in absolute value). Models with higher-level domains generally

exhibited better fit compared to those without such domains. Given the high correlation among composite scores, using these SEM-constructed scores is statistically equivalent to employing the overall score derived from summing up all itemized scores.

Results from the Tripod data were consistent with those from CLASS. Pairwise correlations of overall composite scores, domain summative scores, and SEM-constructed scores from all items displayed perfect correlation (> 0.99). The only discrepancy was that the second-order structure failed to converge in Tripod data. However, SEMs with domain structures still demonstrated better model fit than those without domain structures.

Several technical nuances are worth noting. Firstly, the SEMs incorporating cross-instrument factors displayed better fit than those only accounting for within-instrument item clustering. However, SEMs with factors solely based on itemized scores, regardless of whether there is cross-instrument inclusion, consistently showed poor fit with the data, notably less optimal than the SEMs using composite scores. Moreover, upon closer examination, the item clusters comprising the cross-instrument factors lacked theoretical justification. Secondly, removing items with insubstantial loadings, such as EI in MQI, marginally enhanced fit, albeit to an insignificant extent. However, in some cases, it is necessary to exclude them to achieve convergence. Lastly, the second-order SEM structure failed to converge when using MQI composite scores instead of MQI itemized scores along with CLASS and Tripod composite scores, possibly due to low internal consistency of MQI in the current dataset. All the aforementioned empirical evidence contributes to the selection of the final SEM with the current structure.

Based on the findings from CLASS, Tripod, and exploratory factor analysis of MQI, the final SEM for constructing instructional quality measures utilizing information from all three

observational protocols entails a second-order model comprising one general factor (InsQ), a CLASS composite score, a Tripod composite score, and items from MQI (Figure 3. 1). Results and basic fit indices for elementary data are shown in Figure 3. 2, with additional model fit indices in Appendix Table A3. 4. A similar pattern emerged in secondary school data, albeit with a slightly less optimal model fit, with results shown in Figure 3. 3 and additional model fit indices in Appendix Table A3. 5.

Notably, although the SEMs with the aforementioned bi-factor structure achieved excellent model fit, particularly for the elementary school data, the analytic results do not support the construction of a general index for instructional quality incorporating the information across the three instruments. This is primarily due to the minimal loading of MQI onto the proposed overall second-order factor, *InsQ*, which is 0.0024 in the elementary data and 0.0054 in the middle school data. Based on current evidence, subsequent analyses in this project will examine the causal impacts of CKT on each instrument individually, with a particular focus on MQI. Future research should employ sophisticated methods, such as item factor analysis based on item response theory, to account for measurement errors for each instrument and further explore the possibilities of a comprehensive framework and a general factor for instructional quality. Nevertheless, the analytic results presented in this session provide extensive evidence on the underlying structure within and across instruments, validating the strategy of analyzing fine-grained yet distinctive aspects of instructional quality.

2. Causal analysis of CKT effect on instructional quality

(1) Methodological framework

a. Definition of the causal effect

To define the causal effect of CKT on instructional quality in terms of potential outcomes (e.g. Huber et al., 2020; Imai & Van Dyk, 2004; Imbens, 2000), let Z denote the CKT level of a teacher, which is a continuous treatment variable with support $\mathcal{Z}$. Suppose $z \; and \; z'$ are two possible CKT levels, $z \neq z' \; and \; z, z' \in \mathcal{Z}$. Unlike a binary treatment variable, the binary indication of treatment or control group does not apply here. Instead, $Z = z'$ represents a treatment value different from z and can be called an alternative treatment status.

Let $M(z)$ denote the potential instructional quality if a class has been assigned to treatment z, i.e. $M(z)$ and $M(z')$ represent the potential levels of instructional quality if a class has been assigned to treatment values z or z', respectively. Both are functions of the treatment. In the later causal mediation analysis where instructional quality serves as the mediator in the CKT effect on learning outcome, $M(z)$ and $M(z')$ will be the potential intermediate outcomes.

The causal effect of CKT on instructional quality is defined as:

$$\beta^M = \frac{M(z') - M(z)}{z' - z}$$

As treatment variable Z is continuous, suppose β^M is a continuous function of z, which is denoted as $\beta^M(z)$. Define the marginal causal effect of CKT at z as:

$d\beta^M(z) = \frac{M(z+dz)-M(z)}{dz}$ where dz is an infinitesimal change from z.

The marginal causal effect at z can be viewed as the instantaneous rate of change of a continuous treatment effect at value z. In this study, the treatment effect is assumed to be constant across the population of classrooms.

With the multisite randomization design, let subscripts j, m and k represent classroom j within randomization block r in school k. Denote $M_{jrk}(z)$ as the instructional quality of classroom j within randomization block r in school k being taught by a teacher of CKT level at z, the classroom-specific causal effect is:

$$\beta_{jrk}^{M} = \frac{M_{jrk}(z') - M_{jrk}(z)}{z' - z}$$

It should be noted that classrooms are made exchangeable at the grade level within schools by the experimental design. In other words, systematic between-class variation in student composition within a school has been removed by design. This can be viewed as a matched pair design given that each classroom was matched with another classroom at the same school comparable in student composition; they differ only in the treatment value.

In a study of a binary treatment, the population average causal effect is $E\left[\beta_{jrk}^{M}\right]$. However, with a continuous treatment condition, in general, the causal effect for the population may be a continuous function that allows for heterogeneity across values of z. Here, the population of interest is the population of classrooms. For simplicity, in the primary analyses, I assume that the causal effects are constant across the classrooms in the population and conducted additional analyses that included a between-school random component for these effects. However, it should be noted that the causal effects may be not only heterogeneous but also potentially non-linear, in which case the analytic model would be misspecified. I empirically inspect the data to detect nonlinear patterns.

b. Identification assumptions

Assumption 1. Stable Unit Treatment Value Assumption (SUTVA)

The unit of treatment assignment is classroom in this study. The potential outcomes for all classrooms are assumed to meet the SUTVA assumption within schools. It assumes that the

potential outcomes for one classroom won't be affected by the teacher CKT in other classrooms within the same school after adjusting for the clustering within randomization blocks and within schools. This assumption will be violated if teachers and students from different classrooms at the same school have frequent interact, such as through teacher collaborations by sharing course materials or coordinating teaching plans to synchronize progress within the school. If this assumption is violated, the causal effects of CKT on instructional quality and student achievement will be biased.

Assumption 2. Ignorable treatment assignment

Let vector X_{jrk} denote teacher characteristics and G_{jrk} denote class composition that might have different values between classes within a school. Given covariates X_{jrk} and G_{jrk}, treatment assignment Z_{jrk} is assumed to be independent of potential outcomes $M_{jrk}(z)$ for classroom j within randomization block m in school k.

$$M_{jrk}(z) \perp Z_{jrk} \mid X_{jrk} = x, G_{jrk} = g$$

By the design of the randomization, student composition of the participating classes was made exchangeable within the randomization block at the same grade level within schools. i.e., G should be independent of Z within randomization blocks if the randomization was not contaminated. However, X is likely associated with Z and could be a confounder of the Z-M relationship given that CKT is one of the many teacher characteristics that are randomized to classrooms. Teachers who have different levels of CKT may also differ in other characteristics such as advanced degrees, years of teaching experience, racial identity, and other background characteristics. These teacher characteristics might be associated with both CKT and instructional quality and confound their relationship. However, if these teacher characteristics

would influence instructional quality only through affecting CKT, in econometric research, such variables are referred to as 'bad controls' and should therefore be excluded from the analysis.

Another source of the confounding effects might come from noncompliance and non-random attrition, which might make the classrooms not exchangeable, i.e., classroom characteristics systematically associated with CKT. Both types of potential confounders can be and will be empirically tested.

Additionally, as the CKT assessment was administered to teachers throughout year 2, in some cases, it was observed possibly after classroom videos had been recorded for observational ratings. Nonetheless, it seems reasonable to assume that, in most cases, a teacher's CKT was relatively stable during the two-year study period.

(2) Analytic strategy

a. Model specifications.

The analytic model employed in the causal analysis is a three-level hierarchical linear model (HLM) with multi-level random intercepts and school-level random treatment effects to accommodate the clustering of exchangeable classrooms within randomization blocks at grade-by-school level.

To address variations in students' developmental stages and educational environments, the data were segmented by school grade levels, and parallel analyses were conducted for each sub-sample separately. This approach simplifies the model specification. Initially, the subsamples have only a limited number of grades present: 2 in the elementary-school sample, 3 in the middle-school sample, and 1 in the high-school sample. I include an indicator variable for each grade in elementary-school and middle-school samples. Additionally, given that the dataset comprises only six districts, I included indicators for each district to capture fixed differences

between the districts. Further model specifications, such as grade-by-treatment interactions and/or a random-slope structure, were explored within segmented sub-samples. However, their inclusion was found to diminish the goodness-of-fit of the analytic models.

Let subscript *jrk* indicates classroom *j* in randomization block *r* of school *k.* Independent variable CKT is centered at the mean within randomization blocks. Coefficients π, θ, β represents classroom-level coefficients, randomization block-level coefficients and school-level coefficients respectively. $Grade_{rk}$ represents a vector of indicator variables for grade level, e.g., in middle school sample, $Grade_{ij} = [Grade5_{rk}, Grade6_{rk}]^T$ where $Grade5_{rk}$ and $Grade6_{rk}$ are binary variables, indicating the classroom j in randomization block r is at 5^{th} grade level. $District_k$ represents a vector of indicator variables for districts, $District_{0j} = [District2_k, \ldots, District5_k]^T$, where $District2_k$ to $District5_k$ are binary variables. Residual variance components e, ε, d represent variation at the classroom-, randomization block- and school-levels and are assumed to follow zero mean normal distributions with variance σ_e^2, σ_ε^2, σ_d^2 respectively. Therefore, the following model specifications represent optimal fit according to the current analytic approach for MQI. For CLASS and Tripod, simply replace the term MQI with corresponding terms.

Level-1 Model (Classroom-level)

$$MQI_{jrk} = \pi_{0rk} + \pi_{1rk} CKT_{jrk} + e_{jrk}$$

$$e_{jrk} \sim N(0, \sigma_e^2)$$

Level-2 Model (Randomization block-level)

$$\pi_{0rk} = \theta_{00k} + \theta_{01k} Grade_{rk} + \varepsilon_{rk}$$

$$\pi_{1rk} = \theta_{10k}$$

$$\varepsilon_{rk} \sim N(0, \sigma_\varepsilon^2)$$

Level-3 Model (School-level)

$$\theta_{00k} = \beta_{000} + \beta_{010}\,\text{District}_k + d_{00k}$$

$$\theta_{01k} = \beta_{010}$$

$$\theta_{10k} = \beta_{100} + d_{10k}$$

$$d_{00k} \sim N(0, \sigma_{d0}^2)$$

$$d_{10k} \sim N(0, \sigma_{d1}^2)$$

$$\text{Cov}(d_{00k}, d_{10k}) = \tau$$

Mixed Model:

$$\text{MQI}_{jrk} = \beta_{000} + (\beta_{100} + d_{10k})\,\text{CKT}_{jrk} + \beta_{010}\,\text{District}_k + \beta_{010}\text{Grade}_{rk} + d_{00k} + \varepsilon_{rk} + e_{jrk}$$

To be consistent with the multi-site randomized design, I employed a strategy known as group mean-centering, whereby the independent variable at level 1, CKT, is centered at the mean within each randomization block. In this context, randomization blocks are clusters comprised by same-grade classrooms with exchangeable student composition. Empirically, this centering strategy has proven effective. For instance, upon examining whether classroom composition still confounds the relationship between CKT and instructional quality due to possible contamination of the experimental design, notable findings emerged. Prior to centering, the percentage of minority students in elementary school classrooms significantly predicted CKT latent scores, as well as scores of CLASS and MQI. However, after centering, no classroom composition indicators significantly predicted CKT latent scores, nor did they significantly predict the outcomes, the SEM-constructed factors of instructional quality.

Likewise, the centering strategy also provides evidence that CKT is the primary predictor of instructional quality, rather than other teacher characteristics. By mean-centering both CKT

latent scores and teacher characteristics indicators, certain teacher characteristics e.g., teaching experience in elementary schools and middle schools, demonstrate significant associations with CKT latent scores (results summarized in Table 3. 5). However, only CKT significantly predicts the SEM-constructed factors of instructional quality, which will be discussed in detail in the subsequent section.

(3) Analytic results of causal effects

The outcome variables in the causal analysis are an SEM-constructed factor *MQI* (Italicized font indicates SEM-constructed factor and will be used throughout the following paragraphs), a composite score of CLASS, and a composite score for Tripod. Respectively, I investigated the causal relationships between CKT and each outcome variable within subsamples segmented by school level. It should be noted that the current analysis did not discuss measurement errors in the aforementioned instructional quality factors, as they are the outcome variables of primary interest. While measurement errors in outcome variables increase estimation uncertainty, they do not introduce bias into the estimation of relevant coefficients. However, this may become a concern when these instructional quality factors are predictors in the regression, such as when they act as mediators in a causal mediation relationship.

In elementary and middle school mathematics classrooms, results of causal analyses indicate significant impacts of CKT on *MQI,* while no significant effects were found in high school classrooms (Table 3. 6). Specifically, a one-unit increase, approximately 1 standard deviation (SD) in CKT latent scores significantly predicts a 0.025 (0.179 SD) increase in *MQI* across elementary-school classrooms and a 0.021 (0.231 SD) increase across middle-school classrooms. The corresponding effect sizes, i.e., standardized regression coefficients, which will be used throughout the rest of this study, are 0.154 and 0.212. In contrast to significant CKT

effects on MQI in lower school grades, in 9th grade classrooms, higher levels of CKT do not significantly impact any of the measures of mathematical quality.

Notably, in the analytic model specified above, effects of CKT are allowed to vary by school membership (i.e., random slopes for CKT), where the estimation for random component in the slope provides evidence of between-school variation in CKT impacts. One possible circumstance for between-school variation in CKT impacts is that the effects of hiring a high CKT teacher might be more pronounced in schools with scarce teaching resources compared to those with adequate resources. However, from Table 3. 6, the evidence indicates minimal between-school variation in CKT impacts on MQI. Note that Models including covariance for school-level random components cannot provide accurate standard errors for random-effects parameters using Stata/MP 18.0. Therefore, the covariance has been fixed to zero. This suggests that the benefits of high CKT are relatively consistent across different school environments under this experimental design, and the variation in instructional quality (as measured by MQI) due to CKT might not be fully captured with the current analytic strategy.

With individual domain scores of MQI available, I further explored the causal relationships between CKT and these fine-grained MQI domain scores. This investigation was prompted for following reasons: first, the factor analysis results in prior sessions suggested that MQI lacks strong internal consistency, indicating that causal relationships might differ across individual domains within the MQI instrument. Second, understanding the relationship between more granular dimensions of instructional quality and CKT might offer valuable insights for providing detailed feedback to teachers, a strategy with great practical use. Additionally, results will reveal whether conclusions remain consistent when directly using the original domain scores instead of SEM-constructed scores.

Upon closer examination of the dimension scores within the MQI instrument (results summarized in Table 3. 7), it becomes evident that higher levels of CKT are linked to an increase in the richness of mathematical content, particularly in lower grades. Specifically, a one-unit (approximately 1 SD) increase in CKT latent scores predicts a 0.061 increase in Mathematical Richness ratings in elementary-school classrooms and a 0.049 increase in middle-school classrooms. The corresponding effect sizes are 0.202 and 0.243. Higher CKT also predicts less errors and imprecisions in middle schools, consistent with previous studies (e.g., Hill et al., 2008). Further differentiation between elementary and middle school results reveals that in middle school classrooms, higher CKT predicts a greater degree of student participation in meaning-making and reasoning, whereas in elementary school classrooms, teachers with higher CKT tend to engage more actively with students and mathematical concepts. Higher CKT also significantly predicts a 0.066 decrease in the frequency of errors and imprecision in middle-school classrooms (effect size: 0.243). These results align with previous research (e.g., Hill et al., 2008) and highlight the varying impact of being taught by a high CKT teacher across different school grades. Notably, no significant effects were observed on any MQI dimension in 9th-grade classrooms, consistent with the absence of significant effects on the composite score at this grade level.

In addition to MQI, no significant causal impacts were found for CLASS or Tripod across all school levels (results summarized in Table A3. 6). This suggests that teacher-student interaction and student-perceived classroom experiences, as measured by CLASS and Tripod, capture distinct aspects of instructional quality that may not be directly related to a teacher's math CKT. These findings indicate that while a teacher's CKT is crucial for effective math instruction, it does not necessarily translate into better outcomes in areas assessed by CLASS and

Tripod, such as emotional support, classroom organization, or student engagement. This distinction underscores the multifaceted nature of instructional quality and the need to consider a broad range of factors when evaluating teaching effectiveness and quality of instruction.

It is important to exercise caution when interpreting the significance levels of these estimated regression coefficients. To address the issue of inflated type 1 error rate related to multiple testing, an adjusted criterion for p-values was provided along with summarized results.

VI. Conclusion

This study aimed to investigate the causal relationship between teachers' content knowledge for teaching (CKT) and the quality of their instructional practices in mathematics classrooms. Teacher CKT primarily affects student learning by influencing the quality of teaching strategies and practices in the classroom, where major teaching and learning activities occur. Therefore, understanding CKT impacts on instructional practices become crucial for uncovering the mechanism through which various levels of CKT affect student learning. However, there is limited causal evidence on the relationship between CKT and instructional quality, particularly regarding the impact of CKT on various instruments of instructional quality. This gap is due to limitations such as the lack of experimental data and methodological challenges in measuring instructional quality's complex structure. This research addresses these gaps by: (1) investigating the underlying structures within and across multiple instruments—both subject-general and math-specific observational ratings, as well as student perception survey, and (2) analyzing experimental data from the Measures of Effective Teaching (MET) project. Specifically, it addressed two research questions: (1) Do mathematical teachers with higher levels of CKT deliver instruction with higher quality than those with lower levels of CKT? (2)

How are various instruments of instructional quality causally impacted by variations in mathematics teachers' CKT?

This study examined the underlying structure of three commonly used instruments—Classroom Assessment Scoring System (CLASS), the Mathematical Quality of Instruction (MQI), and the Tripod survey. The analysis revealed a more parsimonious structure than the theoretical frameworks suggested in the original protocols within these instruments. However, despite excellent model fit in SEMs with this bi-factor structure, the evidence does not support combining the three instruments due to the minimal loading of *MQI* (0.0024 in elementary data and 0.0054 in secondary school data) onto the general second-order index for general instructional quality. Advanced methods including item factor analysis based on item response theory should be employed to address measurement errors within each instrument and further explore a comprehensive framework for instructional quality. The results validate the approach of individually analyzing detailed yet distinct aspects of instructional quality.

The causal analysis in this study employed a three-level hierarchical linear model with a random-effect structure to account for the clustering of exchangeable classrooms within randomization blocks within schools—a unique experimental design of the MET project. The analysis revealed significant impacts of CKT on the SEM-constructed factor *MQI* in elementary and middle school mathematics classrooms, particularly reflected by a significant increase in the richness of mathematical content. In contrast to *MQI*, no significant causal relationships were found between mathematical teachers' CKT and teacher-student interaction measured by CLASS and student perception ratings Tripod across all school levels, indicating that these instruments measure distinct aspects of teaching that are not directly influenced by a teacher's mathematical content knowledge. This divergence underscores the multifaceted nature of instructional quality.

Given the current experimental design that randomly assigns exchangeable classes to teachers with different levels of CKT, the analytic results have implications for hiring and assigning high CKT teachers to classrooms. Notably, this study does not evaluate the causal effect of training teachers to improve their CKT on instructional quality, which would require a different type of experimental study that randomly assigns teachers to either receive targeted CKT training or not. Current analytic results suggest that hiring teachers with varying levels of CKT can lead to changes in various dimensions of instructional quality, and these changes differ by grade level. For instance, hiring a high CKT teacher may effectively enhance the mathematical quality of instruction in elementary or middle schools. However, the same strategy may not yield significant improvements in high school settings. School and district administrators must thoroughly understand the specific needs of each grade level and customize their hiring and teacher assignment strategies to target relevant dimensions of instructional quality.

The study presented in this chapter has several limitations. The first limitation lies in the data. The experimental design only spans one year, rendering the experimental data cross-sectional. This limitation means that when comparing CKT effects on instructional quality by grade or across segmented samples, cohort differences cannot be ruled out. Additionally, the high school sample only includes 9th-grade classrooms, limiting the sample size and the generalization of the results to other high school grades.

The second limitation stems from the assumptions regarding the distribution of treatment effects. For simplification, the causal framework of this study assumes a constant causal effect across classrooms within grade levels. However, the treatment effects can be heterogeneous and even non-linear, depending on where the teacher's CKT stands on the ability spectrum. Future

analyses should relax the assumption of constant treatment effects and explore various types of heterogeneity and non-linearity in these effects.

Third, compared with summative scores or simpler structures, this study uncovered a parsimonious bifactor structure across the three commonly used instruments that yields a great fit for the current data. However, current empirical evidence did not support the construction of a general instructional quality index, particularly with relatively low internal consistency regarding MQI. Methodologically, combining multiple instruments that measure distinct theoretical domains of instruction and represent different perspectives requires intricate design and careful consideration. This is particularly relevant because the Tripod instrument is rated by students with year-long classroom experience, whereas other instruments are rated by trained professionals based on limited-time classroom observations. Advanced methods, such as item factor analysis based on item response theory, can be employed in future research to account for measurement errors within each instrument and further explore the possibilities of a comprehensive framework and a general index for instructional quality.

Following this, considering that the ultimate goal of instruction is to improve students' learning outcomes, the causal evidence on the relationship between CKT and instructional quality should be further utilized to identify how CKT affects student learning, whether through direct pathways or specific instructional practices. These questions will be addressed in the next chapter.

Chapter Four. Does Instructional Quality Mediate the Impacts of CKT on Student Achievement in Mathematics: Evidence from A Casual Mediation Analysis

I. Introduction

Content knowledge for teaching (CKT) is presumed to have a significant influence on students' academic outcomes. The theoretical underpinning of this presumption lies in the belief that teachers with deeper CKT are better equipped to deliver high-quality instruction, thereby facilitate student learning and improving their academic performance (Baumert et al., 2010; Campbell et al., 2014; Charalambous, 2010; Hill et al., 2005, 2011; Kelcey, 2011; Metzler & Woessmann, 2012; Rockoff et al., 2011). Notably, the effective transmission of teacher content knowledge relies heavily on the quality of instructional practice. This suggests that instructional quality may serve as a mediator in the relationship between CKT and learning outcomes. Yet prior research on these relationships has yielded inconclusive results regarding the potential mediation mechanism, which can be attributed to several factors. First, measures of CKT and instructional quality are inconsistent across studies, with each research team designing their own measures or choosing certain items from an existing instrument based on their specific focuses, making it difficult to integrate and generalize the findings. Second, causal evidence is lacking, particularly those employing a mediation framework to analyze relations among the three constructs. Experimental designs targeting teacher CKT in previous studies have typically been based on professional development programs of short duration. These studies often failed to detect significant improvements because CKT is difficult to change in a short time frame, limiting their ability to observe impacts on instructional quality and student achievement. These limitations underscore the need for more rigorous and comprehensive research to clarify the relationships between CKT, instructional quality, and student outcomes.

To fill in the gaps of previous research, the study presented in this chapter primarily answers the following research question: How does the quality of instruction, particularly mathematical quality of instruction, mediate the impact of mathematics teachers' CKT on student learning?

By utilizing experimental data from the Measures of Effective Teaching (MET) project, where between-class contextual differences are removed by design within each randomization block at the school-by-grade level, this study seeks to provide more robust evidence on the causal relationships among CKT, instructional quality, and student outcomes. This is an extension of previous chapters where I explored natural variation of CKT and the causal relationship between CKT and instructional quality.

The analytic results of this study enhanced our understanding of how CKT influences student outcomes. Specifically, the total effect of a one-unit (approximately 2 SD) increase in CKT from the randomization block mean at the grade-by-school level resulted in a 0.017 total increase in student test scores in elementary schools, a 0.037 total increase in middle schools, and a 0.0002 total increase in high schools, with effect size being 0.035, 0.090 and 0.0004. Further decomposing the total effects into mediating pathways through mathematical quality of instruction (MQI) ratings, reveals compelling findings. The indirect causal effects through the SEM-constructed factor *MQI* were 0.003, 0.008 and -0.025, with effect sizes being 0.002, 0.006 and -0.025 respectively. These findings further elucidate the mediation mechanism of CKT on student achievement through the mediating pathways of mathematical instructional quality and enhance our understanding of how CKT causally influences student outcomes. Moreover, it identifies areas for future research, including potential non-linearity in the causal relationships and the heterogeneity in treatment effects across student subpopulations.

The subsequent sections of this chapter are organized as follows: Section II reviews the evidence from prior research and identify research gaps in studies of relations among CKT, instructional quality, and learning outcomes. Section III presents hypotheses derived from theoretical framework and prior knowledge. The next two sections provide a description of the data and methodology. Section VI presents analytic results, followed by conclusions.

II. Literature Review

This section will commence with a review of existing research evidence concerning the relationships among CKT, instructional quality, and learning outcomes. At the conclusion of the section, a summary of the research gaps will be presented, along with a discussion of how this study positions itself within the field.

1. Evidence on relationship between CKT and learning outcomes.

Prior studies have generally found a positive and significant association between teachers' content knowledge for teaching (CKT) and students' learning outcomes (Baumert et al., 2010; Campbell et al., 2014; Charalambous, 2010; Hill et al., 2005, 2011; Kelcey, 2011; Metzler & Woessmann, 2012; Rockoff et al., 2011). However, some argue that this association may be spurious if higher-achieving students are more likely to be taught by teachers with greater knowledge. This could occur if parents influence school decisions to place their children in classes with better teachers or if administrators assign higher-achieving students to well-qualified teachers to boost performance ratings. An experimental design, particularly one that randomly assigns students to teachers, can eliminate the confounding effect of non-random student sorting.

Experimental evidence on the relationship between CKT and student outcomes primarily comes from professional development (PD) programs where teachers are randomly chosen to receive PD aimed at increasing their CKT. The impact of an improvement in a teacher's CKT on

students' academic achievements is only observable when these PD programs effectively increase CKT (e.g. Carpenter et al., 1989; Jacobs et al., 2007; Perry & Lewis, 2011). In contrast, PD programs that have no or limited effects on improving teachers' CKT and instructional practices fail to detect positive learning gains for students (Garet et al., 2011; Jacob et al., 2017; Roschelle et al., 2010; Santagata et al., 2010). PD interventions typically consist of training programs during the summer break and/or a few meetings during the academic year. Given this limited intensity and duration, short-term changes in teachers' CKT can be minimal or largely unobserved.

When studying PD interventions targeting teacher CKT, prior researchers have found that natural variations in CKT levels among teachers correlate with students' academic achievements, highlighting the potential for future analyses on the relationship between CKT and learning gains (Garet et al., 2011; Santagata et al., 2010). However, PD programs are often limited by the scope of participant recruitment, typically involving teachers from the same school or schools within the same district. Researchers have acknowledged that this restricted recruitment can result in a lack of sufficient heterogeneity in CKT among participants, making it challenging to draw meaningful inferences (e.g. Jacob et al., 2017; Jacobs et al., 2007; Santagata et al., 2010).

In essence, prior research suggests that an experimental design that removes the confounding effects of contextual differences is needed for better revealing the causal impacts of CKT on student achievement. Moreover, considering that CKT is difficult to improve in the short term, it may be more practical to utilize the natural variation in CKT levels to detect its effects, rather than relying on PD programs targeting CKT, which are often limited by uncertainty about whether the interventions can induce sufficient changes in CKT to observe its actual impact. Therefore, a broad range of teacher CKT is necessary, implying the necessity for a large sample

size. The experimental data from the MET project not only encompass a wide range of teacher CKT, but were collected from a unique experimental design that assign comparable classes of students to teachers with varying CKT levels. This design avoids the limitations of relying on short-term PD training improvements in CKT to detect its causal impacts.

Moreover, while the experimental design of prior research and that of the MET project both aim to analyze the causal impacts of CKT, they address different policy questions. An experimental design that randomly assigns teachers to PD programs investigates whether improving a teacher's CKT would impact instructional quality and student learning. In contrast, an experimental design that randomly assigns students to teachers who naturally vary in their CKT levels, as seen in the MET project, examines whether students would benefit from being taught by teachers with higher CKT. The answers to these two questions are not necessarily the same and have different practical implications. Even if a specific PD program fails to improve teacher CKT, schools may nonetheless choose to recruit and retain teachers with high CKT if they prove to be better teachers. Therefore, causal evidence on the mediated impacts on student achievement of being taught by teachers with relatively high CKT, derived from analyzing the experimental data from the MET project, can provide important practical implications to improve instructional quality and student achievement through recruiting or retaining teacher with high CKT, offering a different perspective from prior research.

2. Relation between instructional quality and student outcomes.

Conceptually, researchers agree that higher quality instruction can lead to positive student outcomes. However, reaching a generalizable conclusion about the relationship between instructional quality and student outcomes is challenging. Previous studies are limited by evidence generated using inconsistent assessment of instructional quality that are designed or

selected based on the researchers' specific focuses and not generalizable across studies, instructional quality measures that did not incorporate multifaceted constructs, and in research settings that cannot ensure causality. Given that the relationship between instructional quality and student outcomes is not the primary focus of this chapter, I will briefly review the scholarship on how instructional quality generally relates to student learning outcomes, particularly highlighting evidence and limitations that are relevant to the effect pathway of CKT.

Research on the relation between instruction practice and student achievement originates from the process-product studies of teaching. In their seminal paper, Brophy and Good critically reviewed 33 process-product studies, concluding that certain teacher behaviors (summarized in Appendix Appendix Figure

Figure A4. 1) were found related to changes in students' learning outcomes. It should be noted that as Brophy and Good emphasized, the effectiveness of instructional practices is heavily dependent on the contexts in which instruction actually takes place. Without considering the actual contexts, any attempts of identifying "the most productive" teachers behaviors are futile (Brophy & Good, 1984).

Recent empirical evidence on the correlation between instructional practice and academic outcomes are mixed. While some researchers found instructional quality meaningfully "predicted" student achievement growth (Hill et al., 2011; Hill & Chin, 2018; Kane et al., 2011; Kane & Staiger, 2012; Milanowski, 2004; Tyler et al., 2010), others found the correlation weak and insignificant (Blazar, 2015; Gencturk, 2012; Shechtman et al., 2010).

It is challenging to exhaustively discuss all the reasons contributing to the divergence in research evidence, and such a discussion is beyond the scope of this study. However, a few key factors are worth mentioning. First, the inconsistency in measures makes comparisons across

studies difficult. Prior research focusing on instructional quality has highlighted this issue (Charalambous, 2020; Mu et al., 2022). In most cases, researchers relied on measurements designed by their own teams or adapted from a single observation protocol. Simplified measures might fail to capture the full range of elements that reflect the effects of instructional quality. Second, the variability in research contexts also complicates the ability to draw general conclusions (Lynch et al., 2017). While evidence gathered from case studies and natural associations is extremely valuable and informative, these settings often lack the control needed to ensure causality. Consequently, contextual differences regarding the characteristics of students, teachers, schools, and districts can confound the focal relationship, making it difficult to draw robust conclusions from non-causal evidence.

3. Mediation studies of CKT, instructional quality, and student learning

Several studies have employed the mediation framework to investigate the impacts of CKT on learning outcomes through instructional quality. Evidence from these studies suggests that instructional quality significantly mediates the effects of teachers' CKT on students' learning (Baumert et al., 2010; Kelcey et al., 2019; Kersting et al., 2012).

Baumert and colleagues (2010) found that CKT was a decisive factor in instructional quality related to cognitive challenge levels, instructional support and classroom management. Moreover, CKT has predictive power for student progress. Their study revealed a significant and positive effect of CKT on students' learning outcomes, mediated through changes in certain aspects of teachers' instructional practices, such as modifying the cognitive levels of tasks and providing individualized learning support (Baumert et al., 2010).

Kersting and colleagues (2012) developed a measure called the Classroom Video Analysis (CVA) to assess teacher CKT, which requires teachers to view and discuss classroom

video clips in written responses. Specifically, the CVA encompasses four domains: "mathematical content," "student thinking," "suggestions for improvement," and "depth of interpretation." When analyzing pairwise associations among the three constructs—CKT, instructional quality, and learning gains—significant associations were found between CKT (as measured by CVA) and instructional quality, as well as between instructional quality and learning. However, no significant associations were found between CKT and learning gains, except for the "suggestions for improvement" domain. Based on these findings on pairwise correlations, they further adopted a mediation model to focus on the indirect effect of CKT on learning gains through changes in instructional quality. They found that the estimated indirect effects of the three CVA domains other than "suggestions for improvement" were significant, suggesting that different domains of CKT might affect learning gains differently. In this case, the "suggestions for improvement" domain of CKT directly impacted learning gains, while the other three domains influenced students' learning gains primarily through the indirect pathway of changing instructional quality (Kersting et al., 2012).

A recent correlational study by Kelcey and colleagues extended the mediation analyses to a larger sample with more participants observed over time than in prior studies. Although the authors acknowledged that the evidence were tentative since none of the analyses were causal, they confirmed that the correlations among CKT, instruction, and achievement were significant, consistent with prior research. Their results have several important implications: First, knowledge should be proximal to instruction to yield noticeable returns. Thus, returns to teachers' knowledge in a specific subject, such as mathematics, are more likely to be found in subject-specific domains of instructional quality. Second, district contexts matter. Evidence showed that the mediating role of instruction was more pronounced in districts with coherent

instructional guidance, long-term investment in instructional reforms, and cognitively challenging state tests. Third, their results revealed heterogeneity in returns to teacher knowledge on instructional quality across the percentiles of the MQI domain scores. Specifically, teachers who made more observed mistakes benefited more from increases in CKT than teachers with very few observed mistakes. Teachers who scored high in employing ambitious mathematics instruction benefited more from increases in CKT compared to teachers who rarely employed such instruction. In contrast, returns to instructional quality on student achievement were consistent and did not show noticeable heterogeneity across the distribution of student achievement. Fourth, contextual differences at various levels, (i.e., students, classrooms and districts) moderated different effect pathways. Specifically, district contexts, such as having a coherent and sustained system for instructional guidance, moderated the pathway between instructional quality (mediator) and student achievement (outcome) in both magnitude and direction. Districts with coherent and sustained instructional guidance system had higher returns of instructional quality on student outcomes. Classroom contexts, reflected by two MQI sub-domain scores (ambitious mathematics instruction and frequency of mathematical errors occurred in class) primarily moderated the pathway between CKT (treatment) and instruction (mediator). Returns of high CKT were stronger in classrooms that scored higher in ambitious mathematics instruction and had fewer observed mistakes in teaching (Kelcey et al., 2019).

Overall, research evidence on the mediation effects of CKT on student achievement remains limited, despite the crucial role of understanding the mechanisms by which CKT impacts student learning. This gap may stem from challenges in measuring the three constructs, especially CKT and instructional quality, as well as issues related to the availability of comprehensive data.

4. Summary of research gaps and potential contribution of this study

Firstly, it is well-established that experimental designs are ideal for identifying causal links. However, previous research on the relationships among CKT, instructional quality, and student achievement has largely focused on observational studies with minimal controls, primarily examining associations rather than causality. Among the studies that have explored causal links, many relied on PD interventions that may not induce significant changes in CKT in the short term, Hence, it is necessary to utilize the natural variation of CKT from a larger sample size to ensure sufficient heterogeneity for detecting CKT impacts. This study aims to address these limitations by utilizing experimental data from the Measures of Effective Teaching (MET) project. Unlike conventional studies, the MET project involves randomly assigning teachers to exchangeable classrooms in its experimental stage. This design eliminates the confounding effects of class-level contextual differences resulting from non-random student sorting within schools. Furthermore, the MET project includes a large sample size, with 537 teachers from 162 schools across six urban public-school districts in the United States. This expansive scope provides a broad range of teacher CKT levels, crucial for detecting its impacts on student learning. Consequently, the evidence obtained from this study will possess higher levels of internal and external validity compared to prior research.

Secondly, prior research has faced limitations in measuring instructional quality and CKT, often using inconsistent measures across studies, which hinders comparisons and generalization of the conclusions. This study utilizes a well-developed measure of mathematical instructional quality, and an enhanced measure for CKT. These improved measures allow for a more nuanced analysis and generates more accurate evidence of causal mediation relationships that previous studies could not achieve.

Lastly, there is a lack of research evidence on the mediation effects of CKT on student achievement. Prior studies have faced challenges in establishing proper measurements for the three constructs, especially CKT and instructional quality, and have been limited by the availability of comprehensive data. This study extends the line of scholarship by attempting to unveil the causal links between CKT and student learning through the mediation of instructional quality. It aims to explore the multifaceted ways in which CKT influences instructional quality and student learning outcomes, considering both direct and mediated pathways as well as contextual differences across educational levels. This mediation analysis will help to clarify the specific mechanisms through which CKT affects student achievement, offering insights into how mathematical instructional quality serve as a mediator for these effects. This comprehensive approach will contribute to a deeper understanding of the interplay between teacher knowledge, instructional practices, and student outcomes, addressing gaps in the existing literature and providing a stronger empirical foundation for educational policy and practice.

III. Hypotheses

Informed by the causal analysis in the previous chapter and prior mediation research, I propose the following hypotheses:

Hypothesis 1. Indirect effect of CKT on student achievement through changing MQI: Math teachers with higher CKT will demonstrate higher mathematical quality of instruction. This includes fewer errors and imprecisions and a greater ability to work with students to solve math problems, which in turn leads to better student learning outcomes.

Hypothesis 2. Indirect effect of CKT on student achievement through changing student perception measured by Tripod: Math teachers with higher CKT will have more

agency to effectively plan and structure class time. This will result in students having a more positive perception of their classroom experience and will better facilitate their learning.

Hypothesis 3. Direct and Alternative Pathways of CKT effect on student achievement: Math teachers with higher CKT will help students achieve better learning outcomes directly or through mechanisms other than those mentioned above. This may include enhanced teacher-student interactions, innovative teaching methods, or other instructional strategies not explicitly covered by the previous hypotheses.

Hypothesis 4. Variation of CKT effects by school level: Educational contexts vary greatly by school level, e.g., students' developmental stages, curriculum variations, and class structure differences. This suggests that the relationship between CKT and instructional quality, as well as subsequent student outcomes, may vary significantly across elementary, middle, and high school settings.

These hypotheses aim to explore the multifaceted ways in which CKT influences instructional quality and student learning outcomes, considering both direct and mediated pathways as well as contextual differences across educational levels.

IV. Data description

The analytic data were derived from the Measures for Effective Teaching (MET) study, a large-scale two-year teacher evaluation project designed to investigate, identify, and comprehensively measure effective teaching skills and practice. The MET project recruited six urban school districts across the United States, involving 2,741 teachers and approximately 160,000 students in 4^{th}- to 9^{th}- grade classrooms for English Language Arts (ELA), Mathematics, or Biology.

This paper focuses specifically on the 537 mathematics teachers, each responsible for teaching one class in 162 public schools across the six U.S. urban school districts during the Academic Year 2011-2012, the year when the MET project conducted randomized experiments that assigned teachers within a school and a grade level to comparable classes of students. The analytic dataset used in this study is identical to the one constructed for the causal analysis presented in Chapter 3. However, the final analytic sample is restricted to students and classes that had valid Mathematics standardized test scores available from state administrative records.

V. Methodology

This section outlines the methodology employed in the study. First, I laid out a general framework for causal mediation analysis within the current research context, where both mediators and treatment variables are continuous. Guided by this analytic framework, I discussed the identification assumptions necessary for analyzing the experimental data obtained from a multisite study that randomized teachers to comparable classes, featuring a continuous treatment and a continuous mediator at the class level. Finally, I specified a series of hierarchical linear models corresponding to the three-step procedures for causal mediation analysis and discussed the decisions that led to the final model setup.

1. A General framework for causal mediation analysis.

(1) Definition of the causal effect

The purpose of a causal mediation analysis is to decompose the average treatment effect of a continuous treatment variable Z on an outcome Y into a direct effect and indirect effects transmitted through mediator(s) M. As one of the important pathways in the mediation framework is with M(z) as the intermediate outcome, the causal framework presented in the previous chapter regarding the causal effects of CKT on instructional quality still applies here.

The causal effect of CKT on the learning outcome is defined analogously as above. Note that in the descriptive analysis of CKT and in the causal analysis of the CKT effect on instructional quality, the respective dependent variables CKT and instructional quality are measured at the classroom level. However, in the causal mediation analysis, the dependent variable is the learning outcome Y and is measured at the student level. Let $Y(z, M(z))$ denote the potential learning outcome of a student assigned to a teacher with CKT level z; and let $Y(z', M(z'))$ denote the student's potential learning outcome if the class was assigned to a teacher with CKT level z' instead.

The causal effect of being taught by a teacher with CKT level z' versus z on the learning outcome is defined as:

$$\beta = \frac{Y\big(z', M(z')\big) - Y(z, M(z))}{z' - z}.$$

Adapting the definitions of causal effects for a continuous treatment and a continuous mediator from those provided in the glossary in Chapter 9 of Hong (2015), the direct and indirect effects are defined as follows:

Let $Y(z', m)$ denote the potential learning outcome when teacher CKT level is at z' but with a mediator value (i.e. the level of instructional quality) counterfactually set at a fixed mediator value m.

The controlled direct effect of the treatment on the outcome is defined as

$$\beta^{CDZ} = \frac{Y(z', m') - Y(z, m')}{z' - z}.$$

This represents the effect of CKT on a student's learning outcome when ratings of instructional quality (mediator) are held at a fixed value m'.

The controlled direct effect of the mediator on the outcome is defined as

$$\beta^{CDM} = \frac{Y(z,m') - Y(z,m)}{m' - m}.$$

This represents the effect of ratings of instructional quality (mediator) on a student's learning outcome under a fixed value of CKT z.

Let $Y\big(z', M(z)\big)$ denote the potential learning outcome when teacher CKT level is at z' but with a mediator value (i.e. the level of instructional quality) counterfactually set at $M(z)$ associated with the alternative teacher CKT level z. The natural direct effect of the treatment on the outcome is defined as

$$\beta^{ND} = \frac{Y\big(z', M(z)\big) - Y(z, M(z))}{z' - z}.$$

This represents the effect of being assigned to a teacher with CKT level z' versus a teacher with CKT level z on a student's learning outcome should the ratings of instructional quality (mediator) remain unchanged by teacher CKT. Notably, the difference between the controlled direct effect of the treatment on the outcome and the natural direct effect of the treatment on the outcome lies in the value of the mediator. In the controlled direct effect, the value of the mediator is set at m, while in the natural direct effect, the value of the mediator is the potential intermediate outcome of being assigned to a teacher with CKT value z, which is a random variable that can take different values.

The natural indirect effect of the treatment on the outcome when the treatment value is fixed at z' is defined as

$$\beta^{NI} = \frac{Y\big(z', M(z')\big) - Y(z', M(z))}{z' - z}.$$

This represents the effect of a change in the ratings of instructional quality (mediator) induced by a change of CKT from z' to z on a student's learning outcome under the condition that the class is assigned to a teacher with CKT level z'.

The pure indirect effect of the treatment on the outcome when the treatment value is fixed at z' is defined as:

$$\beta^{NI} = \frac{Y\left(z,M\left(z'\right)\right)-Y(z,M(z))}{z'-z}.$$

This represents the effect of a change in the ratings of instructional quality (mediator) induced by a change of CKT from z' to z on a student's learning outcome under the condition that the class is assigned to a teacher with CKT level z.

The total effect is equal to the sum of the natural direct effect and the natural indirect effect:

$$\beta = \beta^{NI} + \beta^{ND}.$$

Unlike the descriptive analysis of CKT and the causal analysis of CKT's effect on instructional quality, the outcome variable in the causal mediation analysis—student learning—is measured at the student level. However, since a teacher's CKT and instructional quality generally affect an entire class of students, it is important to note that both the instructional quality ratings (mediator) and CKT (treatment) are measured at the class level. Therefore, for the causal mediation analysis, the population of interest for the outcome variable—student learning—is the population of students, while the population of interest for the treatment effects on the intermediate outcome (mediators) is the population of classes. The subscripts for Y and M in the definitions of class-specific and student-specific causal effects will reflect these differences.

Let subscripts *i, j, r and k* represent student *i* in classroom *j* within randomization block *r* of school *k*. Denote $M_{jrk}(z)$ as the instructional quality of classroom *j* within randomization block *r* of school *k* being taught by a teacher of CKT level at z. Denote $Y_{ijrk}(z, M(z))$ as the potential learning outcome of student student *i* in classroom *j* within randomization block *r* of school *k* if the class was assigned to a teacher with CKT level z.

The class-specific effect of being assigned to a teacher with CKT level z' versus being assigned to a teacher with CKT level z on the class's instructional quality ratings (mediator):

$$\beta_{jrk}^{M} = \frac{M_{jrk}(z') - M_{jrk}(z)}{z' - z}.$$

The student-specific causal effects including the following:

Total effect:

$$\beta_{ijrk} = \frac{Y_{ijrk}\left(z', M_{jrk}(z')\right) - Y_{ijrk}(z, M_{jrk}(z))}{z' - z}.$$

Natural indirect effect:

$$\beta_{ijrk}^{NI} = \frac{Y_{ijrk}\left(z', M_{jrk}(z')\right) - Y_{ijrk}(z', M_{jrk}(z))}{z' - z}.$$

Natural direct effect:

$$\beta_{ijrk}^{ND} = \frac{Y_{ijrk}\left(z', M_{ij}(z)\right) - Y_{ijrk}(z, M_{jrk}(z))}{z' - z}.$$

(2) Identification assumptions

Assumption 1. Stable Unit Treatment Value Assumption (SUTVA)

The unit of treatment assignment and that of mediator value assignment is classroom for this project. However, the outcome is measured at the student level. When the teacher CKT level and the level of instructional quality are fixed, the average potential outcome for all student in the same classroom and the same school is assumed to be stable after adjusting for the clustering within classrooms and within schools. This assumption will be violated if teachers and students from different classrooms at the same school have frequent interactions with one another, leading to spillover effects between classrooms.

Assumption 2. Ignorable treatment assignment

Treatment assignment Z_{jrk} is assumed to be independent of potential outcomes $Y_{ijrk}(z, m)$ and potential mediators $M_{jrk}(z)$ for all possible values of z and m for classroom j within randomization block r of school k.

$$Y_{ijrk}(z, m), M_{jrk}(z) \perp Z_{jrk}$$

By the design of the randomization, student composition of the participating classrooms was made exchangeable within every randomization block at each school; and teachers were assigned randomly to these classrooms. If the experimental design is implemented well, between-class differences in student composition will be removed and the treatment effects on the mediator and the outcome can be identified. The ignorability assumption will be violated if there exists noncompliance or non-random attrition, which will compromise the exchangeability of classrooms.

Although randomly assigned to classrooms, teachers may differ not only in CKT but also in other characteristics including teacher qualifications and demographic backgrounds. We argue that teacher qualification indicators such as college major or SAT scores in Mathematics may affect instructional quality and student achievement primarily through teacher CKT. Controlling for such teacher qualification measures would be problematic when the current research question focuses on the CKT effects. Therefore, I chose not to control for these strong predictors of CKT.

Furthermore, a teacher's demographic backgrounds may or may not match the demographics of the majority of students in a class. In general, teacher-student demographic match is expected to contribute to instructional quality and student learning. Within a randomization block, suppose that a teacher who has a demographic match with the students has lower CKT than a colleague who does not have such a match. When this is the case, teacher-

student demographic match will be a potential confounder that will lead to an underestimation of the potential benefit of being taught by a higher-CKT teacher. Yet within another randomization block, suppose that a teacher who has a demographic match with the students has higher CKT than a colleague who does not have such a match, in which case the potential benefit of being taught by a higher-CKT teacher will be overestimated. We assume that, when averaging the CKT effects over all the randomization blocks, the negative bias and the positive bias will be cancelled out. A sensitivity analysis can be employed to assess the potential consequences when the assumption is violated.

Assumption 3. Ignorable mediator value assignment

$$Y_{ijrk}(z, m) \perp M_{jrk}(z), M_{jrk}(z')$$

This assumption states that a teacher's instructional quality is independent of the potential outcomes. We assume that it is valid when the randomization design is well implemented. This assumption will be violated if there exists noncompliance or non-random attrition, which will compromise the exchangeability of the classrooms within a randomization block. As before, we assume that teacher qualifications would affect student learning primarily through teacher CKT; we additionally assume that teacher-student demographic match would affect student learning primarily through teacher CKT and instructional quality. Hence there is no need to control for these teacher characteristics.

In addition to the aforementioned identification assumptions, I assume that teacher CKT, which was measured in the previous year, remained mostly unchanged during the two years of study. This assumption is likely valid because previous studies have found that CKT is mostly stable within a short period of time (Garet et al., 2011; Jacob et al., 2017; Roschelle et al., 2010; Santagata et al., 2010).

(3) Model specifications

To analyze the mediation effects, I have specified a 4-level hierarchical linear model with students as the primary unit of analysis with multi-level random intercepts and school-level random treatment effects. These students are clustered within classrooms, which are nested within the randomization blocks that are at the grade-by-school level. The analysis for estimating mediation effects involves three main steps, assuming the treatment-outcome, treatment-mediator, and mediator-outcome relationships are all linear. Firstly, I estimate the total effects of CKT on students' learning outcomes without including mediators (the Z-Y relationship). Secondly, I examine the treatment effects on mediators (the Z-M relationship). This step is equivalent to the analysis conducted in Chapter 3; thus, I use the same analytic model to ensure consistency. Lastly, I investigate the relationship between the outcome and treatment transmitted through the mediator. Empirically, this involves regressing the student outcome Y on Z and M. The indirect effect estimates are calculated utilizing the estimates from these three steps. Additionally, I also investigate the mediated relationship with a treatment-by-mediator interaction. This involves regressing the student outcome Y on Z, M, and the interaction term ZM. The model specification and results can be found in Appendix.

The following decisions concerning the model setup have resulted in the final model specifications. First of all, the major decision relates to the validity of modeling the treatment-mediator, mediator-outcome, and treatment-outcome relationships as linear. To closely examine whether the relationships among CKT, instructional quality ratings, and student achievement are linear, I have created a series of two-way scatterplots by school levels (Figure 4. 1 to Figure 4. 12), including those between CKT and student achievement, between instructional quality factors, and between instructional quality factors and student achievement. The scatterplots

suggest that the relationships are mostly linear. Thus, for the current study, the linearity assumptions are acceptable. Nonetheless, future research may explore the possibilities of fitting non-linear models to examine the relationships among the three constructs mentioned above, especially between CKT and instructional quality ratings.

The second decision concerns the selection of covariates. In the analytic model, covariates such as pre-test scores and indicators of socioeconomic status (SES) are primarily included at the student level. The rationale for including these student-level covariates is twofold. First, it is theoretically important to include students' pre-test scores and SES backgrounds as they might be potential confounders in the causal mediation analysis should be the randomization design be compromised. Second, including student-level covariates that are strong predictors of students' learning outcomes increases the precision of the treatment effect estimation.

Additionally, student-level characteristics such as gender and racial identity typically do not directly predict a teacher's CKT or the overall instructional quality of a class. The analytic model assumes no interactions between student-level covariates and CKT or between student-level covariates and instructional quality factors, as CKT and instructional quality serve as the treatment and mediator, respectively, at the class level. However, if such interactions do exist, the model would be misspecified. Even though a Z by X interaction might exist in predicting M or Y, if X is not a confounder, omitting the interaction will not introduce bias (for a detailed proof, see Appendix in Chapter 3 of Hong, 2015).

I chose not to include covariates such as the proportions of students eligible for free/reduced-price lunch and average student pre-test scores at classroom or school level. First of all, these higher-level covariates are aggregated from student-level data. Pre-test scores, racial

identity and socioeconomic status indicators have already been included at the student level in order to increase precision of regression precision. If the experimental design is well implemented, class-level differences in student composition will be removed, thus, one does not need to include these class-level covariates to get unbiased estimate of treatment effects. Notably, treatment effects might vary by school contexts, however, in this research context, such heterogeneity is negligible, which was confirmed in Chapter 2. Lastly, not including higher-level covariates in the analysis greatly reduces regression complexity. Thus, only student-level model includes controls indicating student backgrounds.

The third decision is regarding centering strategy. To remain consistent with the multisite experimental design, where classrooms within randomization blocks are exchangeable in student composition, and further rule out contextual differences regarding conditions of school SES, I centered the class-level predictors—CKT, mediators, and their interaction term—at the mean value within each randomization block.

The last decision is concerning the inclusion of grade-by-treatment interactions. Considering that samples are segmented by grades, there is not enough grade span within each sub-sample. Any additional grade effects would be accounted for by the grade indicators. Preliminary model comparison results did not show potential model fit improvement by including grade-by-treatment interactions. Henceforth, it is reasonable to assume negligible distinctions in treatment effects between grades within each school level. Consequently, the final model specifications did not include grade-by-treatment interactions.

Let subscripts *ijrk* indicate student *i* in classroom *j* of matched cluster *m* of school *k*. X_{ijrk} indicates a vector of student-level characteristics, including their prior test scores (math scores in AY 2009-2010), eligibility of F/R lunch, racial identities, ELL status, and Special Education

status. $Grade_{rk}$ represents a vector of indicator variables for grade level, e.g., in middle school sample, $Grade_{ij} = [Grade5_{rk}, Grade6_{rk}]^T$ where $Grade_{5rk}$ and $Grade_{6rk}$ are binary variables, indicating the classroom j in randomization block r is at 5th grade level or at 6th grade level. $District_k$ represents a vector of indicator variables for districts, $District_{0j} = [District2_k, \ldots, District5_k]^T$, where $District2_k$ to $District5_k$ are binary variables, indicating to which public-school district the schools belong. Residual variance components e, v, ε, d represent variation at the student-, classroom-, randomization block-, and school-levels, which are assumed to follow zero mean normal distributions with their own variances to be estimated.

Below are the model specifications for the three analytic steps taken to investigate the causal effects of CKT on learning outcomes mediated by MQI. For Tripod, simply replace the term MQI with Tripod.

Step 1. Analyzing the total effect of the treatment on the outcome conditioning on the covariates

Level-1 Model (Student)

$$Y_{ijrk} = \varphi'_{0jrk} + \varphi'_{1jrk}\,\text{pre-test}_{ijrk} + \varphi'_{2jrk}X_{ijrk} + e'_{ijrk}$$

$$e'_{ijrk} \sim N(0, \sigma'^2_e)$$

Level-2 Model (Classroom)

$$\varphi'_{0jrk} = \pi'_{00rk} + \pi'_{01rk}\,CKT_{jrk} + v'_{0jrk}$$

$$\varphi'_{1jrk} = \pi'_{10rk}$$

$$\varphi'_{2jrk} = \pi'_{20rk}$$

$$v'_{0jrk} \sim N(0, \sigma'^2_v)$$

Level-3 Model (Randomization block)

$$\pi'_{00rk} = \theta'_{000k} + \theta'_{001k}\text{Grade}_{rk} + \varepsilon'_{00rk}$$

$$\pi'_{01rk} = \theta'_{010k}$$

$$\pi'_{10rk} = \theta'_{100k}$$

$$\pi'_{20rk} = \theta'_{200k}$$

$$\varepsilon'_{00rk} \sim N(0, \sigma'^2_{\varepsilon})$$

Level-4 Model (School)

$$\theta'_{000k} = \alpha_{0000} + \alpha_{0100}\,\text{District}_k + d'_{000k}$$

$$\theta'_{001k} = \alpha_{0010}$$

$$\theta'_{010k} = \alpha_{0100} + d'_{010k}$$

$$\theta'_{100k} = \alpha_{1000}$$

$$\theta'_{200k} = \alpha_{2000}$$

$$d'_{000k} \sim N(0, \sigma'^2_{d0})$$

$$d'_{010k} \sim N(0, \sigma'^2_{d1})$$

$$\text{Cov}(d'_{000k}, d'_{010k}) = \tau'$$

Mixed Model: $Y_{ijrk} = \alpha_{0000} + (\alpha_{0200} + d'_{010k})\text{CKT}_{jrk} + \alpha_{1000}\,\text{pre-test}_{ijrk} + \alpha_{2000}X_{ijrk}$

$$+ \alpha_{0100}\,\text{District}_k + \alpha_{0010}\text{Grade}_{rk}$$

$$+ d'_{000k} + \varepsilon'_{00rk} + v'_{0jrk} + e'_{ijrk}$$

Step 2. Analyzing the treatment effect on each mediator conditioning on the covariates

Level-1 Model (Classroom-level)

$$MQI_{0jrk} = \pi^{M}_{00rk} + \pi^{M}_{01rk} CKT_{0jrk} + e^{M}_{0jrk}$$

$$e^{M}_{0jrk} \sim N(0, \sigma^{M2}_{e})$$

Level-2 Model (Randomization block-level)

$$\pi^{M}_{00rk} = \theta^{M}_{000k} + \theta^{M}_{001k} Grade_{rk} + \varepsilon^{M}_{00rk}$$

$$\pi^{M}_{01rk} = \theta^{M}_{010k}$$

$$\varepsilon^{M}_{00rk} \sim N(0, \sigma^{M2}_{\varepsilon})$$

Level-3 Model (School-level)

$$\theta^{M}_{000k} = \gamma_{0000} + \gamma_{0001} District_{k} + d^{M}_{000k}$$

$$\theta^{M}_{001k} = \gamma_{0010}$$

$$\theta^{M}_{010k} = \gamma_{0100} + d^{M}_{010k}$$

$$d^{M}_{000k} \sim N(0, \sigma^{M2}_{d0})$$

$$d^{M}_{010k} \sim N(0, \sigma^{M2}_{d1})$$

$$Cov(d^{M}_{000k}, d^{M}_{010k}) = \tau^{M}$$

Mixed Model: $MQI_{jrk} = \gamma_{0000} + (\gamma_{0100} + d^{M}_{010k})CKT_{jrk} + \gamma_{0001} District_{k} + \gamma_{0010} Grade_{rk}$

$$+ d^{M}_{000k} + \varepsilon^{M}_{00rk} + e^{M}_{0jrk}$$

Step 3. Analyzing the mediated effect on the outcome via the mediator conditioning on the covariates

Level-1 Model (Student)

$$Y_{ijrk} = \varphi_{0jrk} + \varphi_{1jrk} pre\text{-}test_{ijrk} + \varphi_{2jrk} X_{ijrk} + e_{ijrk}$$

$$e_{ijrk} \sim N(0, \sigma^2_e)$$

Level-2 Model (Classroom)

$$\varphi_{0jrk} = \pi_{00rk} + \pi_{01rk} MQI_{jrk} + \pi_{02rk} CKT_{jrk} + v_{0jrk}$$

$$\varphi_{1jrk} = \pi_{10rk}$$

$$\varphi_{2jrk} = \pi_{20rk}$$

$$v_{0jrk} \sim N(0, \sigma^2_v)$$

Level-3 Model (Randomization block)

$$\pi_{00rk} = \theta_{000k} + \theta_{001k} Grade_{rk} + \varepsilon_{00rk}$$

$$\pi_{01rk} = \theta_{010k}$$

$$\pi_{02rk} = \theta_{020k}$$

$$\pi_{10rk} = \theta_{100k}$$

$$\pi_{20rk} = \theta_{200k}$$

$$\varepsilon_{00rk} \sim N(0, \sigma^2_\varepsilon)$$

Level-4 Model (School)

$$\theta_{000k} = \beta_{0000} + \beta_{0100} District_k + d_{000k}$$

$$\theta_{001k} = \beta_{0010}$$

$$\theta_{010k} = \beta_{0100} + d_{010k}$$

$$\theta_{020k} = \beta_{0200}$$

$$\theta_{100k} = \beta_{1000}$$

$$\theta_{200k} = \beta_{2000}$$

$$d_{000k} \sim N(0, \sigma^2_{d0})$$

$$d_{010k} \sim N(0, \sigma^2_{d1})$$

$$Cov(d_{000k}, d_{010k}) = \tau$$

$$\text{Mixed Model: } Y_{ijrk} = \beta_{0000} + (\beta_{0100} + d_{000k})CKT_{jrk} + \beta_{0200}MQI_{jrk}$$
$$+ \beta_{1000}\,\text{pre-test}_{ijrk} + \beta_{2000}X_{ijrk}$$
$$+ \beta_{0100}\,District_{k} + \beta_{0010}Grade_{rk}$$
$$+ d_{000k} + \varepsilon_{00rk} + v_{0jrk} + e_{ijrk}$$

VI. Analytic results

1. Main variables in the causal mediation analysis.

Outcome variable. The outcome variable in this causal mediation analysis is the mathematics scores from state standardized tests in 2011. It is important to note that standardized test scores may not always be ideal measures for assessing students' learning outcomes, as indicated by previous research (Brophy & Good, 1984; Lynch et al., 2017). State standardized tests typically emphasize lower-order skills, potentially overlooking the full impact of CKT and instructional quality (Lynch et al., 2017). Nevertheless, this study utilizes these test scores due to substantial missing data in other supplementary assessments available in the MET database, such as scores for SAT, ACT, Balanced Assessment in Mathematics, and Algebra (only available for 9th grade students). Future research endeavors should explore strategies to integrate data from multiple sources and enhance the measurement of student outcomes comprehensively.

Mediators. Theories of pairwise relationships between mathematics teacher CKT and MQI, and between MQI and student achievement, are well developed and supported by research

evidence from observational and correlational studies (e.g., Hill et al., 2005, 2018; Kelcey et al., 2019). Empirical evidence from the causal analysis in Chapter Three also indicated significant causal relations between math CKT and MQI. Henceforth, subsequent analyses primarily focus on investigating the SEM-constructed factor MQI as the mediator in the relationship between CKT and student achievement.

Additionally, considering that student perceptions based on year-long experiences may provide a distinctive perspective of instructional quality that complements observational ratings like MQI, Tripod has also been included as a potential mediator in the causal mediation analysis. Moreover, due to the lack of strong theoretical or empirical evidence supporting teacher-student interaction as a potential mediator in the CKT-student achievement relationship, results for CLASS were not included in the primary discussion but can be provided upon request.

Treatment variable. CKT is the latent scores constructed using item response theory-based model in Chapter 2. Given that CKT is only assessed once during the MET project, I assume that a teacher's instructional quality fluctuated only to a relatively trivial degree from the time it was assessed to the time instruction occurred. If teachers' CKT have changed significantly from the time they took CKT assessment, then the analytic results of this study would be biased.

Descriptive statistics of the variables reflecting the three constructs at the elementary, middle, high school levels can be found in Table 4. 1.

Step 1 Results: Total effects of CKT on student achievement

To estimate the total effect of Content Knowledge for Teaching (CKT) on students' test scores, I directly regressed student test scores on CKT (results summarized in Table 4. 2). Specifically, being taught by a teacher with a CKT level of one-point, i.e., approximately 2 standard deviations (SD), higher than the average CKT level within the randomization block at

the grade-by-school level predicts an increase in student math scores by 0.016 in elementary schools, 0.037 in middle schools, and 0.0002 in high schools. The corresponding effect sizes are 0.011, 0.029 and 0.0002.

It is evident that no significant total effects of CKT on test scores were found across all school levels. Although generally positive in the direction, the magnitudes of the CKT effects were indistinguishable from zero. Conceptually, this finding contradicts common sense as higher CKT teachers, despite receiving higher instructional quality ratings, yielded student scores no different from lower CKT teachers with lower instructional quality ratings.

According to Baron and Kenny (1986), a mediated effect is considered absent if the total effect is estimated to be zero. Based on the current estimation results of total effects, one might conclude that there is no mediation of the CKT impact on students' learning outcomes. However, an important exception should be considered: Even when the total effect is zero, nonzero indirect and direct effects can still exist. For instance, a negative indirect effect and a positive direct effect may sum to a zero total effect. Therefore, even though the Step 1 analysis showed that the total effect is close to zero, further investigation is needed to confirm that this is not a case where the direct and indirect effects are canceling each other out.

Step 2 Results: Effect pathways of CKT on instructional quality

As shown in Table 4. 3, mathematical quality of instruction ratings *MQI* is significantly impacted by CKT in elementary schools. Teachers with CKT one unit (approximately 2 SD) higher than the randomization block mean CKT at the school-by-grade level predict a 0.063 increase in *MQI* in elementary schools, a 0.029 increase in middle schools, and a 0.003 increase in high schools, with effect sizes being 0.010, 0.003 and 0.000. The Tripod ratings are not significantly impacted by CKT across all school levels. Surprisingly, high school teachers with

CKT one unit higher than the randomization-block mean predict a 1.088 decrease in their students' Tripod perception ratings, with an effect size of -15.857, though this result is not statistically significant.

The analytic results of Step 2 regarding CKT impacts on instructional quality ratings presented above is not diverge from the analytic results of previous chapter where I analyzed CKT impacts on class-level instructional quality ratings. Here, analyses employed student-level data which provides greater statistical power in comparison with the causal analysis of the CKT impacts on instructional quality factors presented in the previous chapter that was a class-level analysis. Therefore, conceptually, the estimates here reflect how each individual students' exposure to varying CKT levels affects their exposure to different instructional qualities, while the estimates in the previous chapter reflect how the exposure of a class of students as a whole unit to teacher CKT affect their exposure to different instructional qualities. Future analyses may investigate whether these effects with strategy accounting for class size.

Step 3 Results: Evidence on indirect effects of CKT through changing instructional quality

For Step 3, I have fit a series of model regressing student outcome on CKT and instructional quality factors (results summarized in Table 4. 4). While one might argue that treatment-by-mediator interactions potentially exist, I have excluded these interactions from the primary analysis for two main reasons: First, there is no strong theoretical foundation supporting the existence of treatment-by-mediator interactions. Second, preliminary results from models including these interactions (summarized in Table A4.1) indicate that, in most cases, the treatment-by-mediator interactions are not statistically significant. Additionally, regression

models without treatment-by-mediator interactions generally achieved a better fit. Therefore, I primarily discuss results obtained from models excluding interaction terms.

If linearity assumptions on mediated relationships among CKT, instructional quality and student outcome hold, the natural indirect effect of CKT on student outcomes through the mediator pathway can be calculated as $\alpha_{0200} - \beta_{0100}$ or $\gamma_{100}\beta_{0200}$ (Baron & Kenny, 1986; Hong, 2015; MacKinnon et al., 2007). Utilizing estimates from Table 4. 3 and Table 4. 4, one can derive the natural indirect effects for mediation analyses, presuming linearity assumptions are valid (results presented in Table 4. 5). From Table 4. 5, it is evident that deriving the indirect effects through either taking the differences or multiplying two coefficient estimates from previous steps does cause some discrepancy but not alerting. Here I mainly discuss the differences ($\alpha_{0200} - \beta_{0100}$).

In elementary schools, if students are taught by a teacher with one unit (approximately 1.712 SD) increase of CKT from the randomization block mean at the grade-by-school level, their test scores would increase 0.016 in total, with effect size being 0.011. For Mathematical quality of instruction ratings, *MQI*, the indirect effect is around 0.003 (effect size=0.002).

Similarly, in middle schools, the total effect of being taught by a teacher with a one-unit (1.751 SD) increase of CKT from the randomization block mean at the grade-by-school level is 0.037 in total, with effect size being 0.029. Of the 0.037 total effects of CKT, 0.008 (effect size=0.006) is via the indirect effect via, *MQI*.

In high schools, if students are taught by a teacher with a one-unit (1.495 SD) increase of CKT from the randomization block mean at the grade-by-school level, their test scores would increase 0.0002 in total (the corresponding α_{0200} estimate in Table 4. 2), with effect size being 0.0002. The indirect effect of CKT through *MQI* is negative, while CKT positively impacts *MQI*

(corresponding γ_{100} estimate in Table 4. 3). This pattern suggests a complex interaction where instruction of improved mathematical quality (*MQI*) do not translate into higher student achievement at the high school level. Instead, the increased CKT enhances these MQI ratings, which paradoxically predicted lower student achievement. Additionally, the negative indirect effects of CKT via *MQI* in high schools provides an example of how a non-zero indirect effect and a non-zero direct effect can cancel each other out, resulting in a nearly zero total effect. To illustrate, a one-unit (1.712 SD) increase in teacher CKT improved *MQI* by 0.004, which induce a 0.025 decrease in the total effects of CKT on student achievement. Thus, the direct effect of CKT through unspecified pathways other than *MQI* can be derived as 0.0252, considering the total effect of CKT on student achievement is 0.0002.

In contrast to *MQI*, indirect effects of CKT on student achievement through changing student perception ratings Tripod are indistinguishable from zero across all school levels. Note that, although negatively impacted by higher CKT, Tripod does not have substantial indirect effect (estimate: 0.0001, effect size: 0.000) on student achievement in high schools. This evidence suggests that student perceptions may not mediate the relationship between teacher knowledge and student achievement. Instead, Tripod seems to capture a distinct aspect of instructional effectiveness that is not fully aligned with measures like MQI.

It should be noted that, given that indirect effects are computed as the difference of two coefficient estimates from two separate regressions or the product of two coefficient estimates, the two-stage linear regressions cannot directly provide accurate standard errors for indirect effects, let alone using them for statistical testing, considering that the covariances of the coefficient estimates remain unknown from the current estimation results. Prior researchers have proposed various methods to accurately compute the SEs from two-stage regression estimates,

but they have not reached an agreement on an optimal method. Simulation methods are commonly accepted as a better alternative and should be used in future research (Hong, 2015).

Discussion of the analytic results

Several key considerations must be taken into account when interpreting the current analytic results. First, outcome variables might not be capturing the overall improvement induced by instructional quality. Standardized test scores have long been considered limited being used as outcome variables (Brophy & Good, 1984; Lynch et al., 2017). Importantly, prior research has found that some state tests lack the required cognitive challenge levels to capture the improvement in student learning (Kelcey et al., 2019; Lynch et al., 2017). Furthermore, one should not overlook the possibility that the effects of increased CKT and quality changes in instruction on student learning could be manifested after one year.

Second, there might exist heterogeneity in treatment effects across student subpopulations that are yet to be explored. Certain subpopulations of students, e.g., those living under poverty, from immigrant families, or other disadvantaged backgrounds, might benefit more from an increase in teacher CKT and instructional quality than their more advantaged peers, given that disadvantaged students greatly rely on school education when additional learning support is unaffordable at home. Future research should investigate such heterogeneity in treatment and mediated effects across student subgroups.

Third, the linearity assumption might not hold. Upon inspecting the two-way scatterplots among CKT, instructional quality ratings, and student achievement (Figure 4. 1 to Figure 4. 12), One can suspect the potential existence of non-linearity. In middle-school and high-school subsamples, the relationship between MQI and CKT appears flat when centered CKT scores range between -1 and 1, with linearity becoming more evident at the extreme ends. In the

elementary-school subsample, the relationship between MQI and CKT shows a negative trend when centered CKT values exceed one standard deviation. These non-linearities could be due to random noise, but it is also reasonable to assume that CKT impacts vary based on teachers' positions on the ability spectrum or student performance spectrum, supported by prior correlational studies using mediation frameworks (Kelcey et al., 2019). It is worthwhile for subsequent research to explore fitting non-linear models to investigate the relationships among CKT, instructional quality, and student achievement, especially for that between CKT and mathematical quality of instruction.

Additionally, the mediation analyses suffer from a significant decrease in sample size caused by missing administrative data in mathematics test scores of high-school students from certain states. Examination of missing patterns revealed entire classroom test score absences, resulting in some randomization blocks with only one classroom. Further investigation into this unusual pattern of missing data should be carried out to determine strategies for better utilizing the data in the high-school subsample, or consider using another samples for high schools with additional grade levels and enhanced representativeness.

Methodologically, it should be noted that the conventional multi-step approach based on path analysis using linear regressions employed in this study has its limitations. Notably, a multi-step approach does not provide a direct estimate of the standard errors of indirect effect(s); bootstrapping and alternative testing procedures can be used to ensure the rigor of significance testing (Baron & Kenny, 1986; Hong, 2015; MacKinnon et al., 2007). In future research, rather than relying on the conventional approach, researchers should explore the potential of utilizing novel methods for mediation analyses to overcome the limitations of the multi-step approach. Potential alternative approaches may include but not restricted to structural equation models that

analyzes the mediator model and the outcome model simultaneously and semi-parametric weighting-based causal mediation analysis.

VII. Conclusion

This study was motivated by a common belief that teachers with higher level CKT are better equipped to deliver high-quality instruction, thereby enhancing student learning and improving their academic performance. Central to this theoretical framework is the role of instructional quality as a potential mediator in the relationship between CKT and student outcomes, which also suggested by a few prior studies (Baumert et al., 2010; Kelcey et al., 2019; Kersting et al., 2012). Despite this compelling theoretical framework, prior research has often yielded inconclusive results on the relations among CKT, instruction quality and student achievement, primarily due to methodological limitations such as inconsistent measures of CKT and instructional quality, potential lack of causal validity in the results of mediation analyses, and oversimplified constructs of instructional quality.

To address these gaps, this study intended to answer an important research question: How does CKT affect student learning through the mediation pathways of instructional quality? Unlike previous correlational studies, this research utilized a causal mediation framework to analyze experimental data from the Measures of Effective Teaching (MET) project, which involved randomizing teachers across exchangeable classrooms in six urban school districts. Focusing on 537 mathematics teachers across 162 schools and their 12,204 students, this study employed various measures of instructional quality as well as an improved measure of CKT previously constructed based on item response theory. With these enhanced measures, I conducted a series of analyses to analyze the mediated effects of CKT through various

instructional quality factors individually, aiming to generate fine-grained evidence of causal mediation relationships among CKT, instructional quality and student achievement.

Analytic results revealed compelling findings. In elementary schools, the total effect of a one-unit (1.712 SD) increase in CKT from the randomization block mean at the grade-by-school level resulted in a 0.016 increase in student test scores, with effect size being 0.011. The indirect effect through mathematical quality of instruction (MQI) ratings is approximately 0.003, with effect sizes being 0.002. In middle schools, a one-unit (1.751 SD) increase in CKT led to a 0.037 total increase in student test scores, with effect size being 0.029. The indirect effect through *MQI* is 0.008; the corresponding effect sizes is 0.006. In high schools, the total effect of a one-unit increase (1.495 SD) in CKT on student test scores is 0.0002, with effect size being 0.0002. The indirect effect through *MQI* is negative. The negative indirect effect via *MQI* reveal how non-zero indirect and direct effects can cancel each other out, resulting in an overall total effects that are indistinguishable from zero. For instance, a one-unit increase in CKT improved *MQI* by 0.004, which led to a 0.025 decrease in student achievement. Thus, with a total effect of 0.0002 and an indirect effect of -0.025, the direct effect of CKT through unspecified pathways other than *MQI* was 0.0252.

In contrast to MQI, the indirect effects through student perception ratings, Tripod were indistinguishable from zero across all school level. Given that no significant CKT impacts on Tripod were observed with data at student level, this might suggest that student perception measured by Tripod is not a mediator in the causal relationship between teacher knowledge and student achievement.

However, current study faced challenges of significant reductions in sample size caused by missing administrative data in high-school subsample. Addressing these challenges through

further investigation into missing data patterns is essential for enhancing the reliability and generalization of findings at the high school level.

Furthermore, it is important to highlight several critical aspects when drawing inferences from the findings of this analysis. First, the outcome variables used may not fully capture the overall improvements induced by instructional quality. Second, there may be unexplored heterogeneity in treatment effects across different student subpopulations. Third, the assumption of linearity might not hold, as potential non-linear relationships are suggested by the visual inspection of the original data. Additionally, the analytic results obtained from the high-school subsample are limited by a reduced sample size due to missing data from state administrative records and the inclusion of only one grade span.

Methodologically, the study acknowledges the limitations of the conventional multi-step approach based on path analysis and linear regressions. Future research could benefit from employing simulation studies and alternative testing procedures to strengthen the rigor of significance testing in mediation analyses. Additionally, exploring novel methods such as advanced path analysis using structural equation modeling and alternative weighting methods for mediation analysis could open the opportunities of overcoming the limitations of conventional methods, as well as providing deeper insights into the causal relationships among CKT, instructional quality, and student outcomes.

In conclusion, this study contributes valuable insights into how CKT causally influences student achievement through the mediation pathway of mathematical instructional quality, and the complex nature of the relationships between teacher content knowledge, instructional quality, and student achievement. While direct effects of CKT on student outcomes are minimal, the role of mathematical instructional quality as a mediator warrants further exploration. Future research

should address the limitations identified in this study, including further examining non-linear relationships among the three constructs, investigating heterogeneity of treatment effects across diverse student populations, exploring a better alternative student outcome measure that comprehensively reflect improvement induced by various aspects of instructional quality on student learning, and applying more sophisticated analytical methods such as structural equation models and semi-parametric weighting-based causal mediation analysis. Through these efforts, a deeper understanding of how teacher knowledge impacts student learning can be achieved, ultimately informing educational policy and practice to enhance instructional quality and student outcomes.

Chapter Five. Summary and Future Directions

This concluding chapter synthesizes the key findings from the three papers presented in this dissertation, discusses their implications for educational policy and practice, acknowledges the limitations of the study, and provides recommendations for future research.

By leveraging the rich, longitudinal data from the Measures of Effective Teaching (MET) project, this research has explored the inequitable distribution of teachers' content knowledge for teaching (CKT) observed in six US urban public school districts, the causal impact of CKT on instructional quality, and the mediating role of instructional quality in the causal relationship between CKT and student achievement utilizing a unique experimental design that randomly assign teachers with comparable classrooms. Here I provide an overview of several key findings.

I. Overview of the Key Findings

Study 1: Inequitable Distribution of CKT: Evidence from Natural Variation of The Year 1 Baseline Observational Data

The first study is a descriptive study that innovatively employed item-response-theory-based (IRT) models to generate latent scores of mathematical content knowledge for teaching (CKT). With CKT as a direct measurement for teacher knowledge (reliability=0.82), results of the first study challenged the common reliance on conventional proxies such as advanced degrees and years of experience. The analysis revealed that CKT latent scores are insignificantly, and sometimes negatively, associated with these conventional proxies and value-added scores.

When utilizing CKT to investigate inequitable distribution of teacher knowledge, analytic results uncovered significant systematic inequality in student access to high-quality teachers between schools, particularly affecting high schools with large populations of disadvantaged students. Meanwhile, although substantial within-school variation in CKT exists, it is not

systematically associated with students' prior achievement or socioeconomic backgrounds and is likely due to pure chance in classroom assignments and natural variation of CKT among teachers in current schools, factors unrelated to inequitable allocation of educational resources.

Study 2: Causal Relationship Between CKT and Instructional Quality

The second study investigated the causal effects of CKT on the quality of instructional practices in mathematics classrooms utilizing multiple commonly used instruments for instructional quality, including Classroom Assessment Scoring System (CLASS), Mathematical Quality of Instruction (MQI), and student perception survey Tripod.

The results of factor analyses attempted to uncover a more parsimonious structure than the theoretical frameworks suggested in the original protocols of these instruments for measuring instructional quality. Although the SEMs with a bi-factor structure achieved excellent model fit, particularly for the elementary school data, the empirical evidence does not support the construction of a general higher-level index that can incorporate the information across instruments. Specifically, this is primarily due to the minimal loading of MQI onto the proposed overall second-level factor, *InsQ*, as well as its relatively low internal consistency. Meanwhile, it should also be noted that Tripod instrument is rated by students with year-long classroom experience, while other instruments are rated by trained professionals based on limited-time classroom observations. Methodologically, combining multiple instruments that measure distinct theoretical domains of instruction and represent different perspectives requires intricate design and careful consideration that are beyond the scope of this study.

Investigating the causal effects of CKT on instructional quality measured by each instrument, the analytic results revealed significant impacts of mathematical teachers' CKT on the SEM-constructed factor *MQI* in elementary and middle school mathematics classrooms,

particularly reflected by a significant increase in the richness of mathematical content. In contrast to MQI, no significant causal relationships were observed across all school levels between CKT and teacher-student interaction measured by CLASS or student perceived classroom experience measured by Tripod.

Study 3: Mediation of Instructional Quality in the Relationship Between CKT and Student Achievement

The third study explored the mediation pathways through which CKT affects student achievement via two instruments of instructional quality, MQI and Tripod. The analytic results are compelling.

In elementary schools, a one-unit (1.712 SD) increase in teacher CKT from the randomization block mean led to a 0.016 total increase in student test scores (effect size: 0.011). The indirect effect through mathematical quality of instruction ratings *MQI* was approximately 0.003 (effect sizes: 0.002), while the effect pathways through Tripod were insignificant and minimal (estimate: -0.001, effect size: -0.001). In middle schools, a one-unit (1.751 SD) increase in CKT resulted in a 0.037 total increase in student test scores (effect size: 0.029). The indirect effect through *MQI* was 0.008 (effect sizes: 0.006). The indirect effects via Tripod were indistinguishable from zero and not statistically significant. In high schools, the total effect of a one-unit increase (1.495 SD) in CKT on student test scores was 0.0002 (effect size: 0.0002). The indirect effects through *MQI* were negative, whereas the indirect effect via Tripod was positive but minimal (estimate: 0.0001, effect size: 0.0001). The negative indirect effects of CKT on student achievement via *MQI* suggest that non-zero indirect and direct effects can cancel each other out, resulting in an overall total effect that is indistinguishable from zero.

Overall, while direct effects of CKT on student outcomes appears to be minimal, nuanced causal mediation pathways through instructional quality warrants further exploration. It is evident that the indirect causal effects through mathematical quality factor *MQI* was more pronounced, while the effect pathways of CKT on student achievement through Tripod was negligible across all school levels.

Furthermore, the study highlights several critical aspects readers should consider when drawing inferences from these findings of the causal mediation analyses. First, it should be noted that the outcome variables, standardized test scores provided by state administrators might lack the required cognitive challenge levels to capture the improvement in student learning (Kelcey et al., 2019; Lynch et al., 2017). Second, there may be unexplored heterogeneity in treatment effects across different student subpopulations. The treatment effects might be larger for certain subpopulations of students. For example, those students who cannot afford additional learning outside of school might benefit more from an increase in teacher CKT and instructional quality given that they greatly rely on school education. Third, the analytic results obtained from the high-school subsample are limited by a reduced sample size due to missing data from state administrative records and the inclusion of only one grade span.

. In conclusion, this causal mediation study contributes valuable insights into how CKT causally influences student achievement through the causal mediation pathways of various instructional quality factors, and the complex nature of the relationships between teacher content knowledge, instructional quality, and student achievement. The study also identifies areas for subsequent research, including further examining non-linear relationships among the three constructs, investigating heterogeneity of treatment effects across diverse student populations, exploring a better alternative student outcome measure that can comprehensively reflect

improvement in student learning, and applying more sophisticated analytical methods such as structural equation models and semi-parametric weighting-based causal mediation analysis. Through these efforts, a deeper understanding of how teacher knowledge impacts student learning can be achieved, ultimately informing educational policy and practice to enhance instructional quality, student achievement and ultimately contributes to education equity.

II. Implications for Policy and Practice

The findings of this dissertation underscore the critical need for policies addressing the inequitable distribution of CKT. Ensuring that all students, particularly those in schools with high concentrations of disadvantaged students, have equal access to high-quality teachers is crucial. States and districts should consider developing and implementing equitable hiring practices that prioritize CKT and providing additional resources and support to schools with lower levels of teacher CKT.

Given the significant impact of CKT on instructional quality, especially in elementary and middle school classrooms, professional development programs targeting CKT are essential. Such initiatives should not only focus on enhancing teachers' mathematical content knowledge and teaching skills in the long term, but also facilitate teacher collaboration, coaching, and mentoring to support instructional improvement with current teacher resources.

Policymakers and practitioners should recognize the distinctive impacts of being taught by a high CKT teacher across different school levels, taking into account students' developmental stages and corresponding needs. While it is a common practice to assign high CKT teachers to grades and subjects where they can have the most significant impact, this strategy must be implemented cautiously to avoid controversy. School and district administrators must ensure that their decisions do not impede educational equity and equal access to quality education. When

assigning teachers based on their CKT levels to optimize instructional quality and student outcomes, it is crucial to maintain fairness and inclusivity.

These policy and practice implications aim to create a more equitable educational landscape, ensuring that all students benefit from high-quality instruction and have the opportunity to achieve their full academic potential.

III. Limitations and Future Directions

The studies presented in this project are subject to several limitations. Firstly, the experimental design of the MET project only spans one year, rendering the experimental data cross-sectional. This limitation means that when comparing CKT effects on instructional quality by grade or across segmented samples, cohort differences cannot be ruled out. Additionally, the high school sample only includes 9th-grade classrooms, limiting the sample size and the generalization of the results to other high school grades.

Secondly, the assumptions regarding the distribution of treatment effects can be extended. For simplification, the causal framework of this study assumes a constant causal effect across classrooms made exchangeable by experimental design at the grade-by-school level. However, the treatment effects can be heterogeneous and even non-linear, depending on where the teacher's CKT stands on the ability spectrum and across student performance spectrum, as indicated by a prior association study (Kelcey et al., 2019). Future analyses should relax the assumption of constant treatment effects and explore various types of heterogeneity and non-linearity in CKT effects.

Furthermore, there might exist heterogeneity in treatment effects across student subpopulations that are yet to be explored. Certain subpopulations of students, e.g., those living under poverty, from immigrant families, or other disadvantaged backgrounds, might benefit more

from an increase in teacher CKT and instructional quality than their more advantaged peers, given that disadvantaged students greatly rely on school education when additional learning support is unaffordable at home. Future research should investigate such heterogeneity in treatment and mediated effects across student subgroups.

Current results of causal mediation analyses also reveal the needs for an advanced study on the underlying structure of instructional quality factors, particularly when they act as mediators. It is worthwhile for future research to probe into the mediated pathways via multiple parallel mediators in current research contexts, addressing unsolved questions behind the absence of significant mediation effects of CKT on student test scores.

Lastly, future research could benefit from employing simulation studies and alternative testing procedures to strengthen the rigor of significance testing in mediation analyses. Moreover, exploring novel methods such as advanced path analysis using structural equation modeling and alternative weighting methods for mediation analysis could open the opportunities of overcoming the limitations of conventional multi-step methods, as well as providing deeper insights into the causal relationships among CKT, instructional quality, and student outcomes.

In summary, this dissertation contributes valuable insights into the relationships between CKT, instructional quality, and student outcomes. By addressing critical gaps and employing rigorous analytical approaches, this research provides a robust empirical foundation for future educational policies and practices. The findings underscore the importance of content knowledge for teaching in shaping instructional practices and student achievement and highlight areas for future investigation. Ultimately, this work aims to inform efforts to enhance educational equity and effectiveness, ensuring that all students receive high-quality instruction from knowledgeable and skilled teachers.

Tables

Table 1. 1 Core Elements of the MET study by Year

Elements	**Year 1 (Academic Year 2009-2010)**	**Year 2 (Academic Year 2010-2011)**
1. District Administrative Data		
School	○ Grade, enrollment size, student composition	○ No new data is collected
Teacher	○ Demographics, professional background	○ No new data is collected
Student	○ Prior scores on state test, current test scores, gender, ethnicity, free lunch status, program participation	○ Prior scores on state test, current test scores, gender, ethnicity, free lunch status, program participation
2. Classroom Videos		
For teachers that are generalists (Subject Matter Generalists)	○ Eight video sessions per teacher: each day recorded ELA, Mathematics for four days	○ Eight video sessions per teacher: each day recorded ELA, Mathematics for four days
For specialist teachers (Subject Matter Specialists)	○ Four video sections per teacher: each day recorded two sections for two days	○ Four video sections per teacher: each day recorded one section for four days
3. Classroom Video Scoring		
Subject Matter Generalists	○ Each video scored with CLASS and FFT. Additionally, ELA sessions scored with PLATO, Math sessions scored with MQI	○ Each video scored with CLASS and FFT. Additionally, ELA sessions scored with PLATO, Math sessions scored with MQI
Subject Matter Specialist	○ English: CLASS, FFT, PLATO ○ Math: CLASS, FFT, MQI ○ Biology: QST	○ English: CLASS, FFT, PLATO ○ Math: CLASS, FFT, MQI ○ Biology: QST

Table 1. 1 Continued.

4. Student Outcome Data	○		○	
Grades 4-5	○	SAT-9 open-ended reading; Balanced Assessment in Mathematic; Student Perception (or Tripod) Survey	○	SAT-9 open-ended reading; Balanced Assessment in Mathematics; Student Perception (or Tripod) Survey
Grades 6-8	○	SAT-9 open-ended reading; Balanced Assessment in Mathematics; Student Perception (or Tripod) Survey	○	SAT-9 open-ended reading; Balanced Assessment in Mathematics; Student Perception (or Tripod) Survey
Grade 9	○	ACT Quality Core English Grade 9; ACT Quality Core Algebra I; ACT Quality Core Biology; Student Perception (or Tripod) Survey	○	ACT Quality Core English Grade 9; ACT Quality Core Algebra I; ACT Quality Core Biology; Student Perception (or Tripod) Survey
5. School Personnel Surveys				
Teachers	○	Teacher Working Conditions Survey	○	MET Teacher Survey; Content Knowledge for Teaching Assessment (Math and ELA)
Principals		-	○	MET Principal Survey;

Table 1. 2 Participation of the MET Study by Year, Study Type, and Unit Level

Unit Level	**Year 1 (Academic Year 2009-2010)** Full sample	Core Study Sample	**Year 2 (Academic Year 2010-2011)** Randomized Sample	Full sample
Districts	6	6	6	6
Schools	317	310	284	317
Teacher	2741	2086	1559	2741
Class sections	4497	1909	1379	4497

Note: There exist a number of teachers whose conditions didn't qualify for randomization requirements being observed in Year 2 (i.e. non-randomized sample).

Table 2. 1 Descriptive statistics of participating teachers' characteristics

VARIABLES	Obs	Mean	Std. Dev.	Min	Max
Gender (1=male)	974	.206	.405	0	1
Race					
White (1=yes)	972	.580	.494	0	1
Black (1=yes)	972	.335	.472	0	1
Hispanic (1=yes)	972	.055	.227	0	1
Generalist (1=yes)	1005	.388	.488	0	1
Master's degree or above (1=yes)	738	.401	.490	0	1
Years of experience					
In total	370	9.927	9.018	0	44
In current district	751	6.777	6.489	0	37
Novice teacher (1=yes)	669	.010	.203	-.661	.682
Value-added scores	1005	0.005	.927	-3.104	2.864

Table 2. 2 Associations of CKT latent scores with teacher characteristics

VARIABLES	Regression Coefficients	Standardized Coefficients	Std. Errors (clustered by schools)	p-value
Gender (1=male)	.183*	.121	.076	.016
Race				
White (1=yes)	.575***	.420	.068	.000
Black (1=yes)	-.699***	-.499	.066	.000
Hispanic (1=yes)	.083	.041	.157	.601
Generalist (1=yes)	-.376***	-.273	.084	.000
Master's degree or above (1=yes)	-.345***	-.251	.078	.000
Years of experience				
In total	-.010	-.031	.006	.086
In current district	-.003	-.008	.006	.617
Novice teacher (1=yes)	-.092	-.043	.087	.293
Value-added scores	.355	.161	.191	.064

Note: Here I present results regressing CKT on each one of the teacher characteristics in the first column without any other controls. These regressions were for descriptive purposes without making causal claims. For binary predictors, the regression coefficients were equivalent to t-tests to compare means of two sub-samples having different values of the predictors. *** $p<0.001$, ** $p<0.01$, * $p<0.05$

Table 2. 3 Descriptive statistics of covariates at teacher-level and school-level within segmented sub-samples

VARIABLES	Obs	Mean	Std. Dev.	Min	Max
Elementary Schools					
Teacher-level					
CKT latent scores	394	-.137	.884	-2.16	2.51
2009 Math scores	394	.022	.471	-2.73	1.55
FRL **(%)**	290	.465	.293	0	1
Black+Hispanic **(%)**	394	.715	.289	.03	1
Black **(%)**	394	.433	.250	0	1
Hispanic **(%)**	394	.054	.103	0	1
ELL **(%)**	394	.150	.178	0	1
Special Education **(%)**	393	.085	.093	0	.56
School-level					
CKT latent scores	394	-.137	.884	-2.15	2.51
2009 Math scores	394	.050	.357	-.703	1.09
FRL **(%)**	290	.461	.280	.05	.98
Black+Hispanic **(%)**	394	.713	.278	.09	1
Black **(%)**	394	.434	.348	.004	1
Hispanic **(%)**	394	.245	.237	0	.96
ELL **(%)**	394	.141	.142	0	.73
Special Education **(%)**	393	.095	.045	.02	.35
Middle Schools					
Teacher-level					
CKT latent scores	373	.040	.927	-2.38	2.35
2009 Math scores	373	-.002	.688	-2.24	1.71
FRL **(%)**	315	.634	.301	0	1
Black+Hispanic **(%)**	373	.690	.302	0	1
Black **(%)**	373	.312	.312	0	1
Hispanic **(%)**	373	.350	.303	0	1
ELL **(%)**	373	.137	.166	0	.98
Special Education	373	.099	.161	0	1
School-level					
CKT latent scores	373	.040	.927	-2.38	2.35
2009 Math scores	373	.101	.410	-.82	.95
FRL **(%)**	315	.627	.278	.08	1
Black+Hispanic **(%)**	373	.685	.279	.17	1
Black **(%)**	373	.307	.280	.02	1
Hispanic **(%)**	373	.351	.275	0	.96
ELL **(%)**	373	.137	.111	0	.58
Special Education **(%)**	373	.078	.048	0	.35
High Schools					
Teacher-level					
CKT latent scores	141	.451	.857	-1.69	2.86
2009 Math scores	131	-.289	.477	-1.64	1.27

Table 2. 3 Continued.

FRL **(%)**	123	.623	.234	.07	1
Black+Hispanic **(%)**	141	.798	.210	.18	1
Black **(%)**	141	.431	.306	0	.1
Hispanic **(%)**	141	.350	.269	0	.97
ELL **(%)**	141	.124	.201	0	.83
Special Education **(%)**	141	.053	.107	0	.59
School-level					
CKT latent scores	141	.451	.857	-1.69	2.86
2009 Math scores	131	-.123	.327	-.929	1.09
FRL **(%)**	123	.611	.225	.104	.96
Black+Hispanic **(%)**	141	.764	.207	.34	1
Black **(%)**	141	.415	.289	.01	.1
Hispanic **(%)**	141	.329	.259	0	.941
ELL **(%)**	141	.111	.165	0	.665
Special Education **(%)**	141	.065	.043	.001	.169

Table 2. 4 Model specifications for Hierarchical Generalized Linear Model

Unadjusted model	**Adjusted model**
Level 1 (Item): $\eta_{ijk} = \log\left(\frac{\varphi_{ijk}}{1-\varphi_{ijk}}\right) = \theta_{jk} - \beta_{ijk} I_{ijk}$	Level 1 (Item): $\eta_{ijk} = \log\left(\frac{\varphi_{ijk}}{1-\varphi_{ijk}}\right) = \theta_{jk} - \beta_{ijk} I_{ijk}$
Level 2 (Teacher): $\theta_{jk} = \pi_{0k} + \pi_{1k}\text{Grade}_{jk} + e_{jk}$ $\beta_{ijk} = \pi_{i2k}$ $e_{jk} \sim N(0, \sigma_e{}^2)$	Level 2 (Teacher): $\theta_{jk} = \pi_{0k} + \pi_{1k}\text{Grade}_{jk} + \pi_{2k} X_{jk} + e_{jk}$ $\beta_{ijk} = \pi_{i3k}$ $e_{jk} \sim N(0, \sigma_e{}^2)$
Level 3 (School): $\pi_{0k} = \gamma_{00} + \gamma_{01}\text{District}_{0k} + r_{0k}$ $\pi_{1k} = \gamma_{10}$ $\pi_{i2k} = \gamma_{i20}$ $r_{0k} \sim N(0, \sigma_r{}^2)$	Level 3 (School): $\pi_{0k} = \gamma_{00} + \gamma_{01}\text{District}_{0k} + \gamma_{02} X_{0k} + r_{0k}$ $\pi_{1k} = \gamma_{10}$ $\pi_{2k} = \gamma_{20}$ $\pi_{i3k} = \gamma_{i30}$ $r_{0k} \sim N(0, \sigma_r{}^2)$
Mixed Model $\eta_{ijk} = \log\left(\frac{\varphi_{ijk}}{1-\varphi_{ijk}}\right)$ $= \gamma_{00} + \gamma_{01}\text{District}_{0k} + r_{0k} + \gamma_{10}\text{Grade}_{jk} + e_{jk} - \gamma_{i20} I_{ijk}$	Mixed Model $\eta_{ijk} = \log\left(\frac{\varphi_{ijk}}{1-\varphi_{ijk}}\right)$ $= \gamma_{00} + \gamma_{01}\text{District}_{0k} + \gamma_{02} X_{0k} + r_{0k} + \gamma_{10}\text{Grade}_{jk} + \gamma_{20} X_{jk} + e_{jk} - \gamma_{i30} I_{ijk}$

Note: Here, subscripts i, j, k represent item i of teacher j in school k. Coefficients π, β represents classroom-level coefficients, and school-level coefficients respectively. Grade_{ij} represents a vector of indicator variables for grade level, $\text{Grade}_{ij} = [\text{Grade5}_{ij}, \ldots, \text{Grade9}_{ij}]^T$ where Grade5_{ij} to Grade9_{ij} are dummy variables. Any one of the grade-level dummy indicates the teacher i in school j teaches the 5th to the 9th grades correspondingly; all equal to 0 indicates the teacher teaches the 4th grade. π_{1j} is a vector of coefficients for vector Grade_{ij}. Similarly, District_{0j} represents a vector of indicator variables for district, $\text{District}_{0j} = [\text{District2}_{0j}, \ldots, \text{District6}_{0j}]^T$ where District2_{0j} to District6_{0j} are dummy variables. Any one of district dummy variables equal to 1 indicates school j being in District 2 to 6 correspondingly; all equal to 0 indicates school j being in District 1. β_{01} is a vector of coefficients for vector District_{0j}. Residual variance components e, r represent variation at the classroom-, and school-levels and are assumed to follow zero mean normal distributions with variance $\sigma_e{}^2$, $\sigma_r{}^2$ respectively.

Table 2. 5 Natural variation by grade, by state and variance decomposition of full sample and subsamples of various levels of schools

VARIABLES	(1) Full sample	(2) Elementary school (4th to 5th)	(3) Middle school subsample (6th to 8th)	(4) High school subsample (9th)
Intercept	-.298*	.574***	.492**	4.226***
	(.150)	(.139)	(.152)	(.739)
Grade indicators				
4th grade	baseline	baseline	-	-
5th grade	.070	.062	-	-
	(.077)	(.069)		
6th grade	.796***	-	baseline	-
	(.108)			
7th grade	.88***	-	.044	-
	(.113)		(.105)	
8th grade	1.143***	-	.342**	-
	(.117)		(.109)	
9th grade	1.422***	-	-	-
	(.123)			
District indicators	YES	YES	YES	YES
District 1	baseline	baseline	baseline	baseline
District 2	.325*	.073	-	.789*
	(.158)	(.190)		(.292)
District 3	-.571***	-.782***	-.607**	.173
	(.096)	(.113)	(.192)	(.278)
District 4	.114	-.056	.166	.552
	(.093)	(.112)	(.164)	(.289)
District 5	-.079	-.226*	-.047	.224
	(.085)	(.115)	(.138)	(.225)
District 6	-.265**	-	-.253*	-
	(.121)		(.139)	
Within-school variation of the random intercept	.414	.315	.501	.486
Between-school variation of the random intercept	.059	.028	.063	.084
Level-2 intraclass correlation	.126	.094	.146	.148
Level-3 Intraclass correlation	.016	.010	.016	.022

Table 2. 5 Continued.

Observations				
Number of classrooms	908	394	373	141
Number of schools	267	109	100	68

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Table 2. 6 Variance comparison by different combinations of fixed effects (Full sample)

	No fixed effects	Only grade fixed effects	Only district fixed effects	Both types of fixed effects
Within-school variation	.423	.413	.420	.414
Between-school variation	.275	.121	.182	.059

Table 2. 7 Variance comparison of CKT latent scores across models including different covariates by school levels (With adjustment for district and grade fixed effects)

	Elementary schools	Middle schools	High schools
No covariates			
Within-school variation	.315	.501	.486
Between-school variation (% of total variation between schools)	.028	.063	.084
Prior achievement			
Within-school variation	.311	.511	.475
Between-school variation	.025	.002	.014
*F/R lunch**			
Within-school variation	.269	.515	.376
Between-school variation	.049	.053	.056
Minority			
Within-school variation	.313	.505	.475
Between-school variation	.024	.026	.045

Note: The sample size for regressions including F/R lunch statuses is notably smaller compared with other regressions. This vast difference came from the missing information of students' F/R lunch statuses in the administrative records of an entire school district.

Table 2. 8 Inequality in CKT distribution by prior performance levels

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Average Math scores in 2009 of students taught by the teacher	-.031	.051	-.081
	(.130)	(.101)	(.237)
School average Math scores in 2009	.248*	.606***	.595*
	(.123)	(.114)	(.240)
Intercept	.543***	.414**	4.874***
	(.139)	(.144)	(1.019)
Within-school variation of the random intercept	.311	.511	.475
Between-school variation of the random intercept	.025	.002	.014
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Level-2 intraclass correlation	.093	.135	.130
Level-3 Intraclass correlation	.007	.0004	.004
Number of classrooms	394	373	131
Number of schools	109	100	67

Standard errors in parentheses

**** p<0.001, ** p<0.01, * p<0.05*

Table 2. 9 Inequality in CKT distribution by Free/Reduced-priced Lunch Status

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of F/R eligible students taught by the teacher	-.624	-1.047*	2.379*
	(.526)	(.515)	(.492)
Proportion of F/R eligible students in school	-.286	-.334	-1.655**
	(.217)	(.266)	(.485)
Intercept	.473**	.524**	4.155***
	(.156)	(.160)	(.735)
Within-school variation of the random intercept	.269	.515	.376
Between-school variation of the random intercept	.049	.053	.056
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Level-2 intraclass correlation	.088	.147	.116
Level-3 Intraclass correlation	.014	.014	.015
Number of classrooms	290	315	123
Number of schools	85	87	60

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Note: The sample size for regressions including F/R lunch statuses is notably smaller compared with other regressions. This vast difference came from the missing information of students' F/R lunch statuses in the administrative records of an entire school district.

Table 2. 10 Inequality in CKT distribution by Minority Status

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of minority students taught by the teacher	-.159	-.542	-.065
	(.555)	(.432)	(.743)
Proportion of minority students in School	-.345*	-.881***	-1.511**
	(.169)	(.236)	(.494)
Intercept	.504***	.332*	3.914***
	(.143)	(.153)	(.742)
Within-school variation of the random intercept	.313	.505	.475
Between-school variation of the random intercept	.024	.026	.045
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Level-2 intraclass correlation	.093	.139	.137
Level-3 Intraclass correlation	.007	.007	.012
Number of classrooms	394	373	141
Number of schools	109	100	68

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Table 2. 11 Inequality in CKT distribution by English Language Learner Status

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of ELL students taught by the teacher	.295	-.519	-.339
	(.358)	(.402)	(1.110)
Proportion of ELL students in School	-.440	-.077	-2.775**
	(.333)	(.652)	(.971)
Intercept	.575***	.495**	4.159***
	(.094)	(.154)	(.736)
Within-school variation of the random intercept	.314	.497	.497
Between-school variation of the random intercept	.025	.064	.026
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Level-2 intraclass correlation	.094	.146	.137
Level-3 Intraclass correlation	.007	.017	.007
Number of classrooms	394	373	141
Number of schools	109	100	68

Standard errors in parentheses

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

Table 2. 12 Inequality in CKT distribution by Special Education Status

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of Special Ed students taught by the teacher	.410	-.172	-.336
	(.583)	(.424)	(.934)
Proportion of Special Ed students in School	.768	-2.202	-5.336*
	(.952)	(1.230)	(2.622)
Intercept	.537***	.443	4.300
	(.143)	(.153)	(.738)
Within-school variation of the random intercept	.315	.506	.484
Between-school variation of the random intercept	.026	.050	.061
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Level-2 intraclass correlation	.094	.145	.142
Level-3 Intraclass correlation	.007	.013	.016
Number of classrooms	393	373	141
Number of schools	109	100	68

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Table 3. 1 Pairwise correlations of MQI items

MQI Items	CWCM	EI	ET	RICH	SPMMR	WWSM
Elementary school data						
classroom work connected to mathematics (CWCM)	1					
errors and imprecision (EI)	0.07	1	/			
richness of mathematics (RICH)	0.05	-0.11	/	1		
student participation in meaning making and reasoning (SPMMR)	0.05	-0.01	/	0.31	1	
working with students and mathematics (WWSM)	0.05	-0.16	/	0.31	0.38	1
Secondary school data						
classroom work connected to mathematics (CWCM)	1					
errors and imprecision (EI)	0.14	1				
explicitness and thoroughness (ET)	0.09	-0.1	1			
richness of mathematics (RICH)	0.33	0.08	-0.05	1		
student participation in meaning making and reasoning (SPMMR)	0.09	-0.14	0.11	0.39	1	
working with students and mathematics (WWSM)	0.14	-0.16	-0.02	0.38	0.5	1

Table 3. 2 Factor loadings of MQI items

MQI Items	**Factor 1**
Elementary school data	
(Eigenvalue)	**0.88**
classroom work connected to mathematics (CWCM)	0.08
errors and imprecision (EI)	-0.17
richness of mathematics (RICH)	0.49
student participation in meaning making and reasoning (SPMMR)	0.54
working with students and mathematics (WWSM)	0.57
Secondary school data	**Factor 1**
(Eigenvalue)	**1.26**
classroom work connected to mathematics (CWCM)	0.3
errors and imprecision (EI)	-0.09
explicitness and thoroughness (ET)	0.06
richness of mathematics (RICH)	0.6
student participation in meaning making and reasoning (SPMMR)	0.64
working with students and mathematics (WWSM)	0.63

Table 3. 3 Descriptive statistics of SEM-constructed factors by grade levels

VARIABLES	Mean	Std. Deviation	Range
Elementary schools (Obs=318)			
CKT latent scores	-.12	.86	[-2.15, 2.27]
InsQ	.01	1.96	[-7.06, 6.46]
MQI	.002	.14	[-.27, .49]
CWCM	.96	.11	[.5, 1]
Error and Imprecision	1.26	.27	[1, 2.5]
Richness of Mathematics	1.33	.26	[1, 2.25]
SPMMR	1.24	.28	[1, 3]
WWSM	1.31	.26	[1, 2.13]
CLASS	49.44	4.01	[33.77, 59.66]
Tripod	73.86	8.24	[35.75, 93.00]
Middle schools (Obs=257)			
CKT latent scores	0.04	0.92	[-2.38, 1.94]
InsQ	.04	3.92	[-14.61, 9.70]
MQI	.002	.091	[-.20, .34]
CWCM	.92	.16	[0, 1]
Error and Imprecision	1.22	.25	[1, 2.13]
Explicitness and Thoroughness	2.23	1.76	[1, 9]
Richness of Mathematics	1.20	.23	[1, 2]
SPMMR	1.15	.20	[1, 1.75]
WWSM	1.27	.25	[1, 2.25]
CLASS	111.84	15.50	[23.57, 143.59]
Tripod	45.078	5.190	[28.76, 57.89]
High schools (Obs=75)			
CKT latent scores	.38	.81	[1.45, 2.09]
InsQ	-2.78	3.26	[-10.82, 4.11]
MQI	-.04	.07	[-.16, .28]
Error and Imprecision	1.14	.18	[1, 1.75]
Explicitness and Thoroughness	2.87	3.07	[1, 22]
Richness of Mathematics	1.15	.23	[1, 2]
SPMMR	1.06	.14	[1, 1.67]
WWSM	1.29	.23	[1, 1.75]
CLASS	41.34	4.50	[29.29, 50.03]
Tripod	106.63	16.06	[61.08, 131.15]
Care	-.11	.56	[-1.51, .94]
Confer	.08	.34	[-1.21, .45]
Captivate	-.12	.67	[-1.83, 1.26]
Clarify	-.12	.53	[-1.58, .76]
Consolidate	-.14	.55	[-1.54, .75]
Challenge	-.09	.39	[-1.13, .48]
Classroom management	-.10	.56	[-1.56, 1.09]

Table 3. 4 Pairwise correlations of SEM-constructed factors

VARIABLES	InsQ	*MQI*	CLASS composite score	Tripod composite score
Elementary school data				
InsQ	1			
MQI	0.81	1		
CLASS composite score	0.52	0.91	1	
TRI composite score	0.16	0.28	0.14	1
Secondary school data				
InsQ	1			
MQI	0.7	1		
CLASS composite score	0.61	0.98	1	
TRI composite score	0.25	0.49	0.37	1

Table 3. 5 Associations of CKT latent scores with teacher characteristics

VARIABLES	Regression Coefficients	Std. Errors (clustered by schools)	p-value
Elementary schools (N=98, n=335)			
Gender (1=male)	-.003	.024	.870
Race			
White (1=yes)	.074	.031	**.016**
Black (1=yes)	-.078	.027	**.004**
Hispanic (1=yes)	.003	.017	.876
Other (1=yes)	.001	.008	.892
Master's degree or above (1=yes)	-.027	.037	.472
Years of experience			
In total	-2.044	.974	**.036**
In current district	-.660	.415	.111
Middle schools (N=279, n=84)			
Gender (1=male)	.018	.032	.576
Race			
White (1=yes)	.180	.029	**.000**
Black (1=yes)	-.166	.026	**.000**
Hispanic (1=yes)	-.004	.017	.821
Other (1=yes)	-.011	.010	.290
Master's degree or above (1=yes)	-.0004	.035	.990
Years of experience			
In total	-1.763	.795	**.027**
In current district	-1.637	.584	**.005**
High schools (N=44, n=98)			
Gender (1=male)	.013	.058	.821
Race			
White (1=yes)	.001	.061	.983
Black (1=yes)	.042	.042	.324
Hispanic (1=yes)	.001	.034	.971
Other (1=yes)	-.044	.033	.182
Master's degree or above (1=yes)	-.059	.082	.470
Years of experience			
In total	-.128	1.333	.335
In current district	-1.597	.877	.069

Note: Here I present results regressing CKT on each one of the teacher characteristics in the first column without any other controls. These regressions were for descriptive purposes without making causal claims. For binary predictors, the regression coefficients were equivalent to t-tests to compare means of two sub-samples having different values of the predictors. The p-values in bold indicate statistical significance of at least 0.05 level.

Table 3. 6 Analytic results for CKT impacts on MQI

VARIABLES	Elementary school (4th to 5th)	Middle school subsample (6th to 8th)	High school subsample (9th)
MQI	.025*	.021**	.007
	(.012)	(.008)	(.010)
Intercept	.044*	.042**	-.046**
	(.022)	(.014)	(.018)
Grade indicators:			
4th grade	baseline	-	-
5th grade	-.029	-	-
	(.016)		
6th grade	-	baseline	-
7th grade	-	-.033*	-
		(.015)	
8th grade	-	-.036*	-
		(.015)	
District indicators:			
District 1	baseline	baseline	baseline
District 2	-.060	-	-.004
	(.047)		(.027)
District 3	-.106**	-.027	.028
	(.030)	(.032)	(.032)
District 4	.011	-.032	.024
	(.028)	(.022)	(.031)
District 5	-.046*	-.008	-.009
	(.028)	(.019)	(.022)
District 6	-	-.050*	-
		(.020)	
Random intercept (random block)	.0002	.001	.0001
Random slope (school)	.000	.000	.0003
Random intercept (school)	.003	.001	.0001
Residual	.014	.005	.002
Level-2 intraclass correlation	.175	.294	.336
Level-3 Intraclass correlation	.164	.153	.321
Number of classrooms	273	238	71
Number of randomization blocks	111	114	35
Number of schools	76	74	32

Standard errors in parentheses *** p<0.001, ** p<0.01, * p<0.05

Table 3. 7 Analytic results for CKT impacts on MQI dimension scores

MQI dimension	Regression Coefficients	Standardized Coefficients	Std. Errors	p-value
Elementary schools				
CWCM	.018	.141	.011	.123
Error and Imprecision	-.052	-.166	.029	.070
Richness of Mathematics	.061*	.202	.023	.008
SPMMR	.011	.034	.027	.673
WWSM	.048	.159	.022	.035
Middle schools				
CWCM	.005	.029	.017	. 768
Error and Imprecision	-.066*	-.243	.026	.012
Explicitness and Thoroughness	.096	.050	.311	.761
Richness of Mathematics	.049	.196	.022	.028
SPMMR	.043*	.198	.018	.020
WWSM	.023	.085	.023	.310
High schools				
Error and Imprecision	.009	.041	.042	.823
Explicitness and Thoroughness	-.238	-.063	.627	.705
Richness of Mathematics	.028	.099	.048	.558
SPMMR	-.016	-.093	.023	.491
WWSM	.011	.039	.041	.779

Standard errors in parentheses

*** p<0.0002, ** p<0.002, * p<0.025 (p-value adjustment for 5 tests)

Table 4. 1 Descriptive statistics of main variables by grade levels

VARIABLES	Mean	Std. Deviation	Range
Elementary schools (n=6,265)			
Student test scores	.026	.939	[-3.26, 3.02]
(Centered)	.000	.828	[-3.60, 2.85]
CKT latent scores	-.139	.842	[-2.15, 2.27]
(Centered)	.002	.584	[-1.71, 1.44]
MQI	-.003	.137	[-.27, .49]
(Centered)	.002	.094	[-.27, .34]
Tripod	73.982	8.657	[35.75, 93.00]
(Centered)	.194	5.450	[-25.03,22.32]
Middle schools(n=5,590)			
Student test scores	.125	.919	[-3.02, 2.82]
(Centered)	-.000	.722	[2.75, 2.93]
CKT latent scores	.048	.891	[-2.38, 1.94]
(Centered)	-.001	.571	[-1.84, 1.84]
MQI	.003	.092	[-.20, .34]
(Centered)	.002	.056	[-.17, .21]
Tripod	111.120	16.203	[23.57, 136.20]
(Centered)	-.185	11.222	[-40.79, 40.79]
High schools (n=354)			
Student test scores	-.365	.730	[-2.28, 1.61]
(Centered)	-.000	.676	[-2.18, 1.93]
CKT latent scores	.658	.932	[-1.19, 2.09]
(Centered)	.003	.669	[-1.19, 1.16]
MQI	-.051	.063	[-.16, .06]
(Centered)	-.010	.058	[-.16, .07]
Tripod	99.916	11.962	[64.50, 125.22]
(Centered)	1.254	9.750	[-30.28, 17.43]

Table 4. 2 Analytic Results for Step 1: Treatment Effects on Student Achievement- Total Effects, α 0200

VARIABLES	2011 Mathematics scores
Elementary schools	
CKT	.016
	(.022)
Middle schools	
CKT	.037
	(.021)
High schools	
CKT	.0002
	(.040)

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Table 4. 3 Analytic Results for Step 2: Treatment Effects on Mediators, γ_{0100}

VARIABLES	*MQI*	Tripod
Elementary schools		
CKT	.063*	.275
	(.027)	(1.391)
Middle schools		
CKT	.029	4.821
	(.020)	(4.145)
High schools		
CKT	0.004	-1.088
	(.020)	(4.662)

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Table 4. 4 Analytic Results for Step 3: Controlled Direct Treatment Effects, β_{0100} and Mediator Effects on Student Achievement given Treatment, β_{0200}

VARIABLES	2011 Mathematics scores (M=*MQI*)	2011 Mathematics scores (M=Tripod)
Elementary schools		
CKT, β_{0100}	.013	.017
	(.022)	(.022)
M, β_{0200}	.137	.0005
	(.141)	(.002)
Intercept, β_{0000}	.117	.116
	(.098)	(.098)
Student-level covariates	YES	YES
District indicators	YES	YES
Random slope (school)	.0003	.000
Random intercept (school)	.134	.135
Random block	.029	.028
Class	.031	.031
Residual	.306	.306
Number of schools	74	74
Number of random blocks	109	109
Number of classrooms	267	267
Middle schools		
Z, β_{0100}	.029	.037
	(.021)	(.020)
M, β_{0200}	.381	.002
	(.221)	(.001)
Intercept, β_{0000}	.300**	.300**
	(.096)	(.097)
Student-level covariates	YES	YES
District indicators	YES	YES
Random slope (school)	.000	.000
Random intercept (school)	.130	.131
Random block	.136	.134
Class	.022	.022
Residual	.266	.266
Number of schools	74	74
Number of random blocks	114	114
Number of classrooms	238	238

Table 4. 4 Continued.

High schools		
Z, β_{0100}	.025	.00008
	(.050)	(.040)
M, β_{0200}	-.846	-.00006
	(.695)	(.003)
Intercept, β_{0000}	-.668***	-.662***
	(.132)	(.139)
Student-level covariates	YES	YES
District indicators	YES	YES
Random slope (school)	.000	.000
Random intercept (school)	.027	.036
Random block	.001	.003
Class	.004	.0001
Residual	.240	.240
Number of schools	14	14
Number of random blocks	15	15
Number of classrooms	26	26

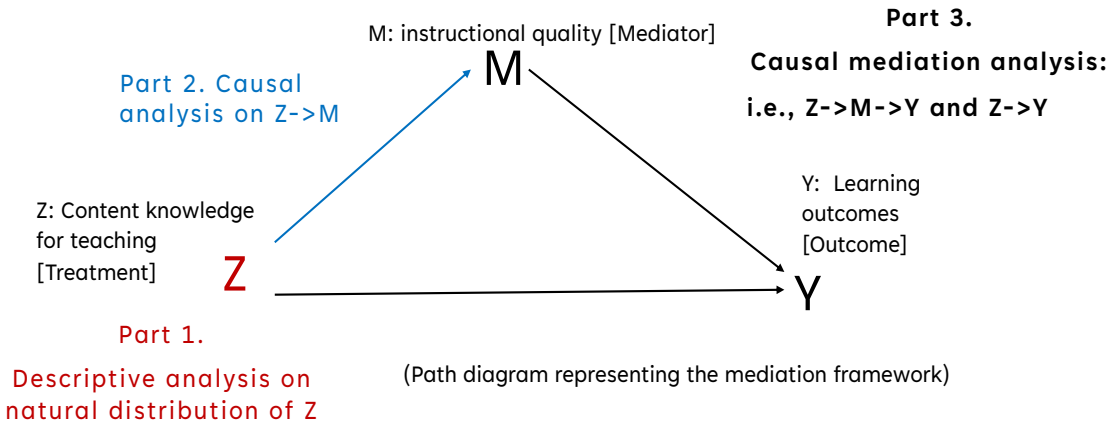
Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Table 4. 5 Indirect effects derived from multi-step regression estimates

Coefficients	*MQI*	Tripod
Elementary schools		
$\alpha_{0200} - \beta_{0100}$	.003	-.001
$\gamma_{100}\beta_{0200}$	.009	.0001
Middle schools		
$\alpha_{0200} - \beta_{0100}$	.008	0
$\gamma_{100}\beta_{0200}$	.011	.010
High schools		
$\alpha_{0200} - \beta_{0100}$	-.025	.0001
$\gamma_{100}\beta_{0200}$	-.003	.0001

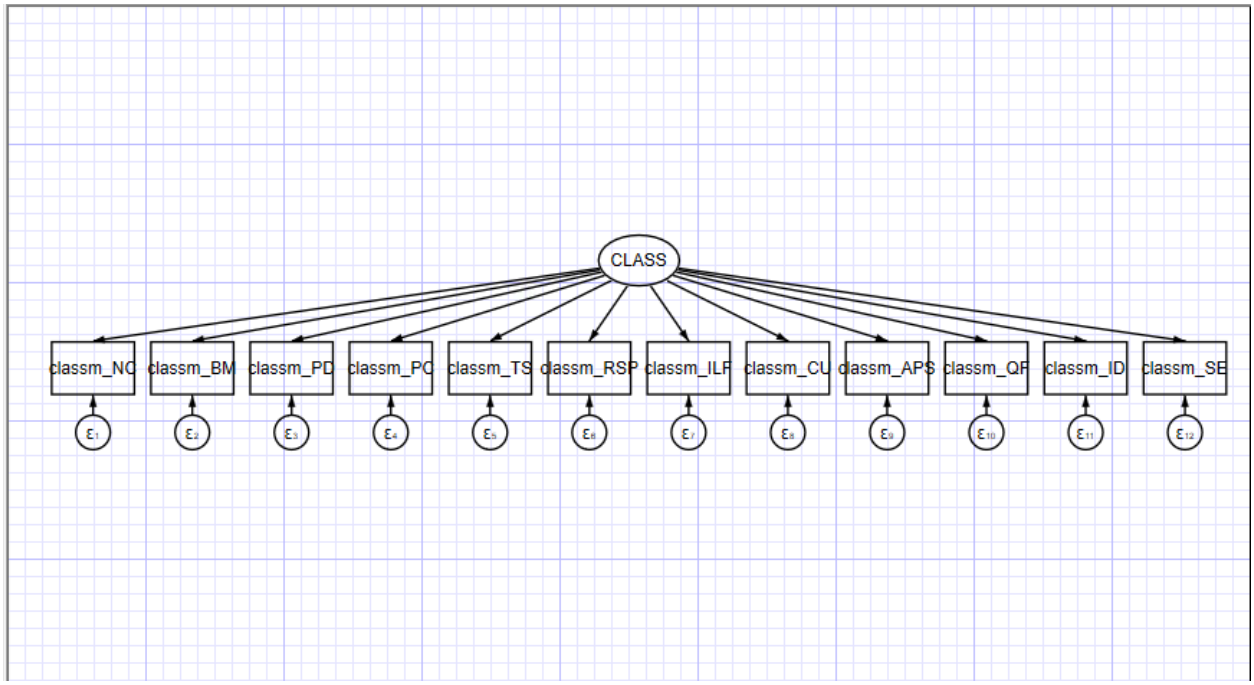
Figures

Figure 1. 1 Structure of dissertation project



(Path diagram representing the mediation framework)

Figure 3. 1 Diagrams indicating structures of three different SEMs for CLASS



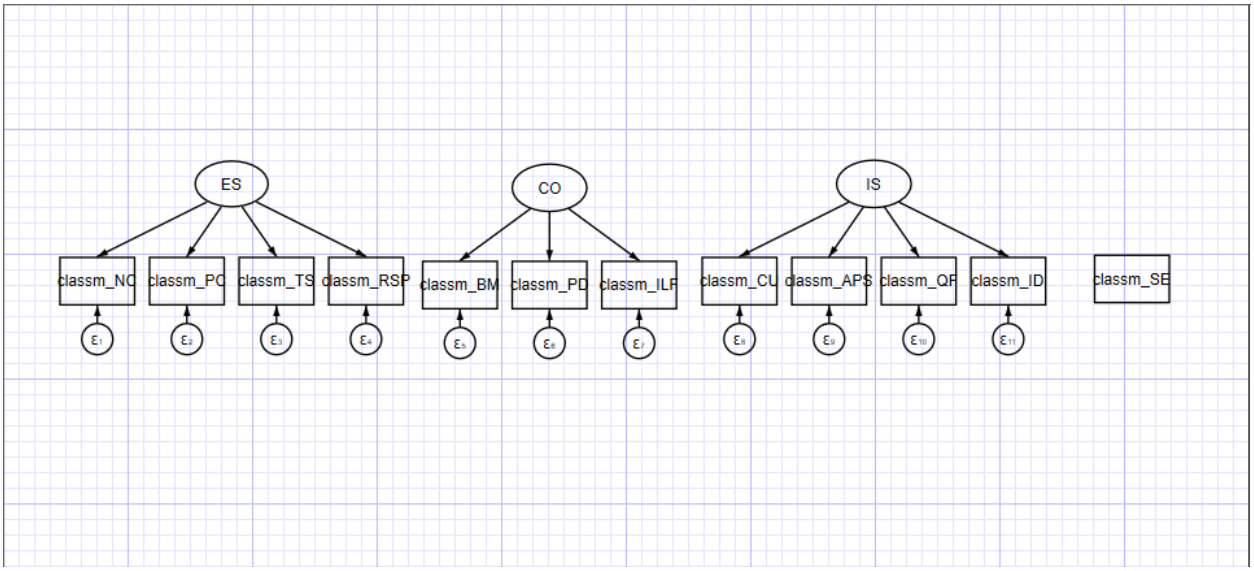


Figure 3.1 Continued.

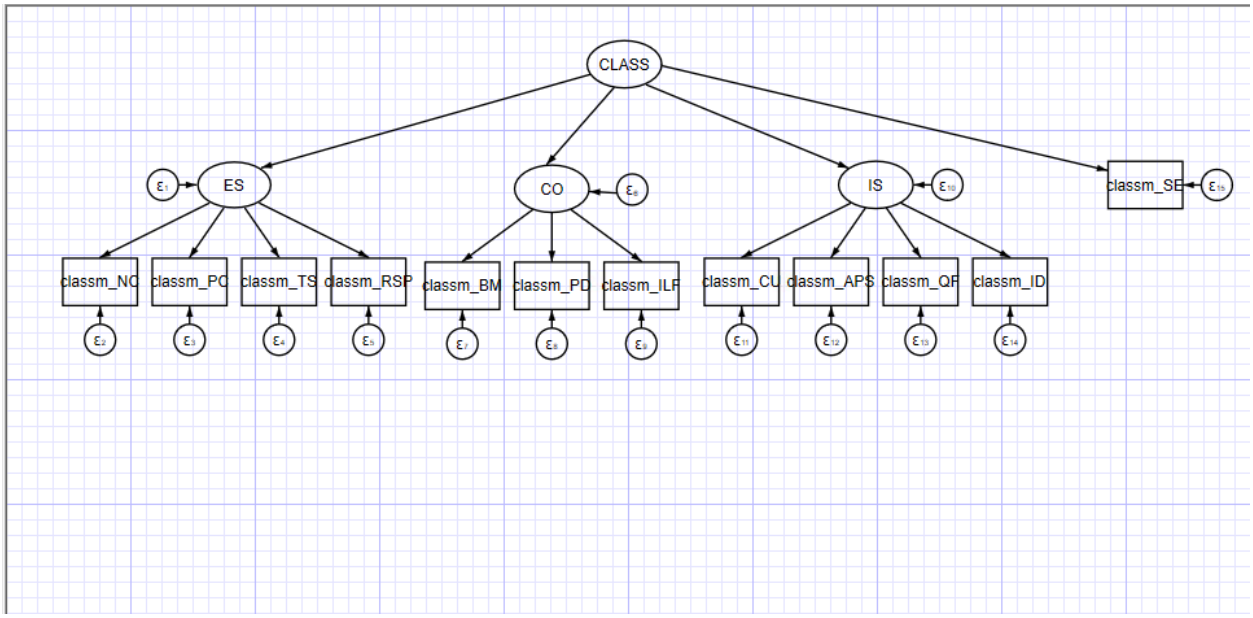
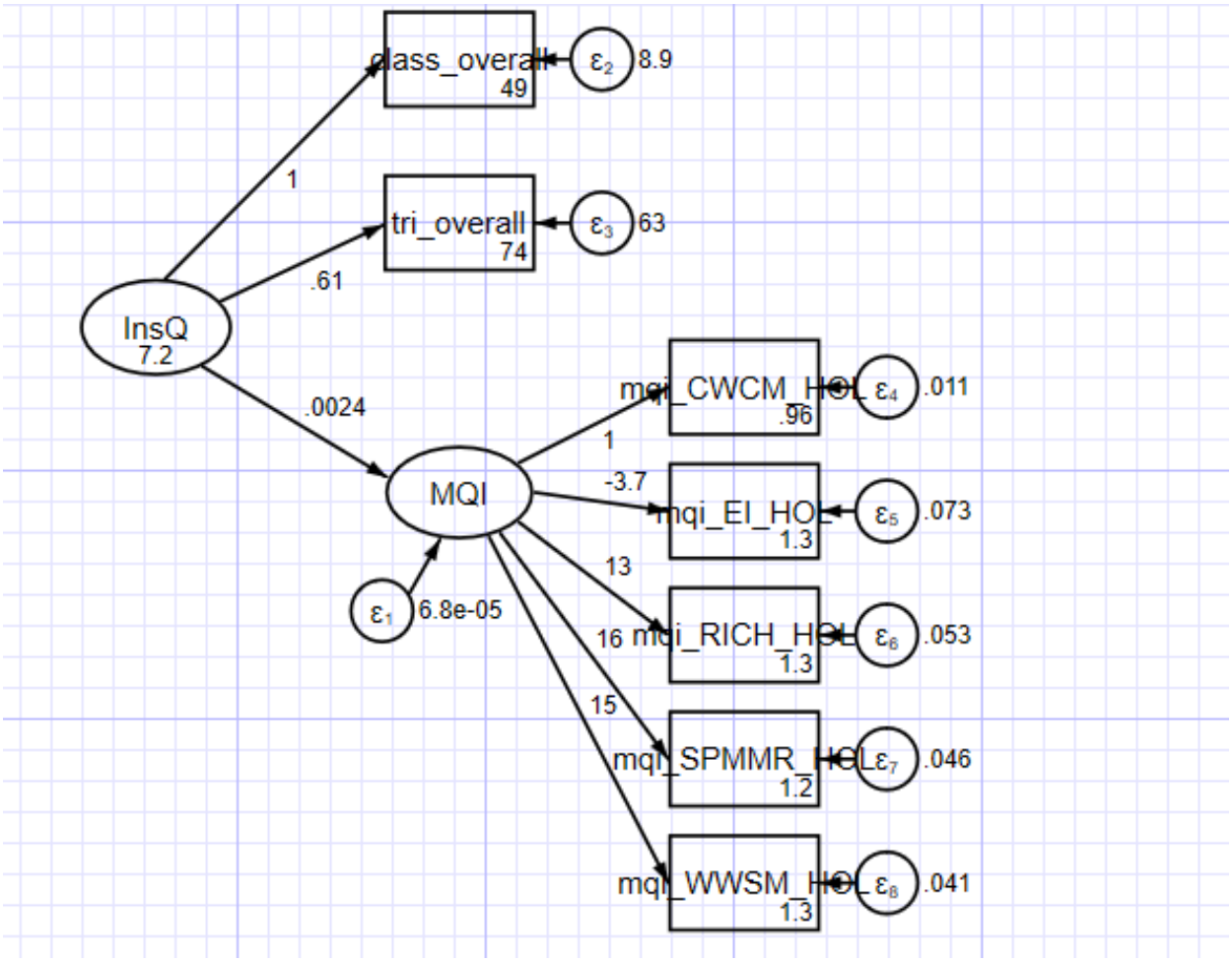


Figure 3. 2 SEM Results of the final structure in path diagram (Elementary data)



Fitness Indices:

ChiSq=14.087

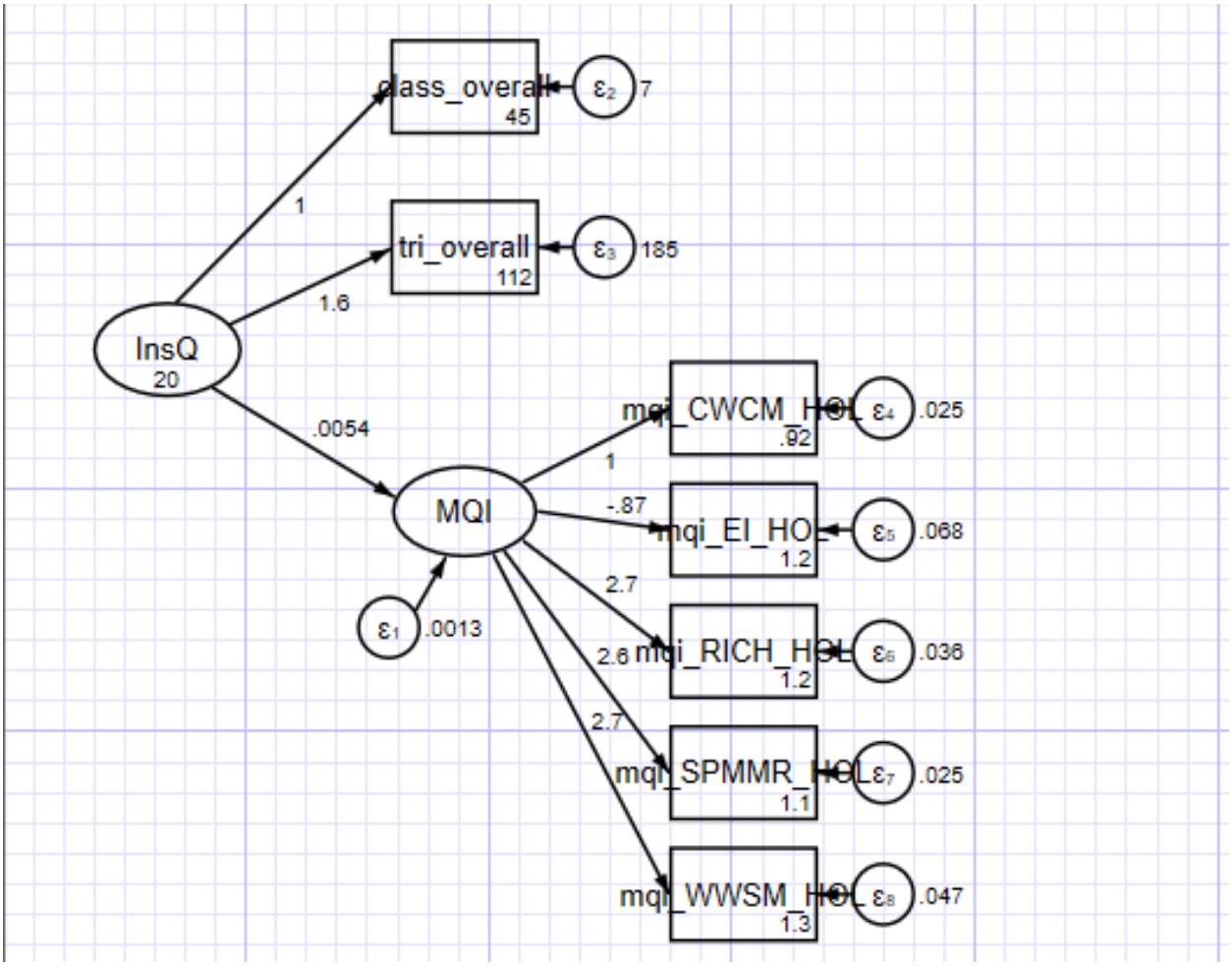
ChiSq/df=1.08

RMSEA= 0.032

CFI= 0.993

TLI= 0.989

Figure 3. 3 SEM Results of the final structure in path diagram (Secondary data)



Fitness Indices:

ChiSq=30.835

ChiSq/df=2.37

RMSEA= 0.072

CFI= 0.889

TLI= 0.821

Figure 4. 1 Pairwise relationship between CKT latent scores (group mean centered) and student math achievement (group mean centered) in elementary-school sample

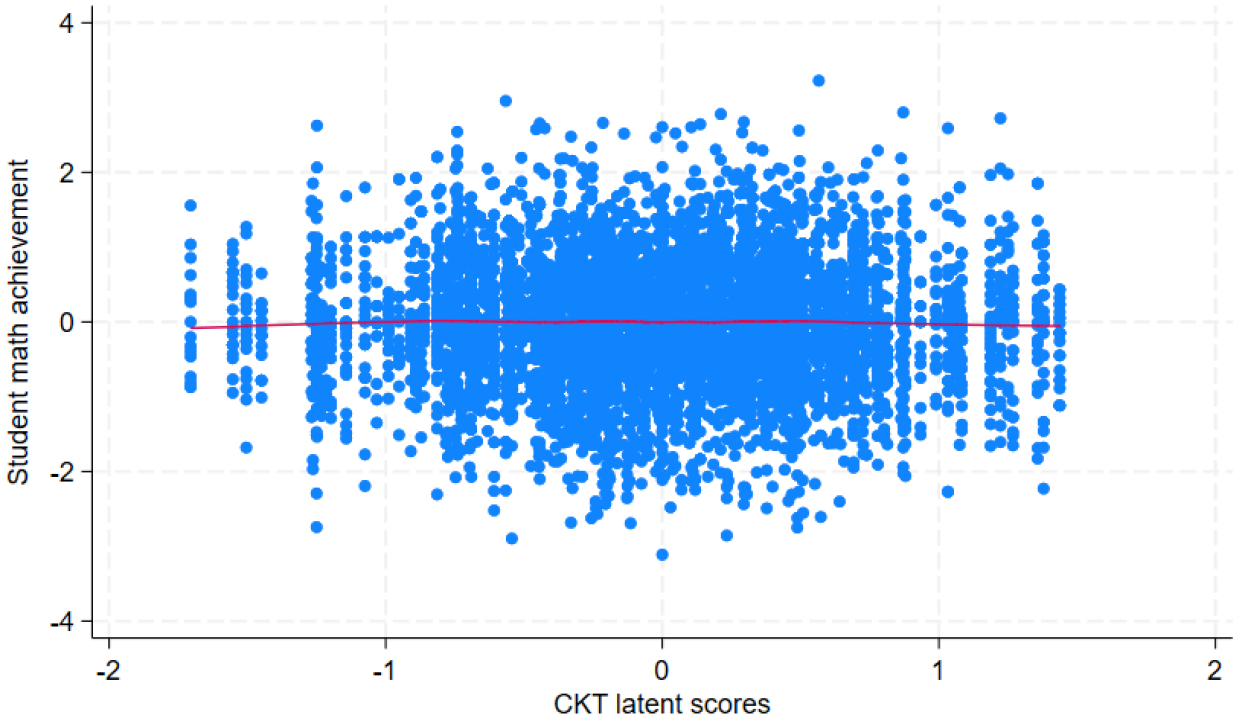


Figure 4. 2 Pairwise relationships between CKT latent scores and MQI, and between MQI and student achievement in elementary-school sample, all variables group mean centered

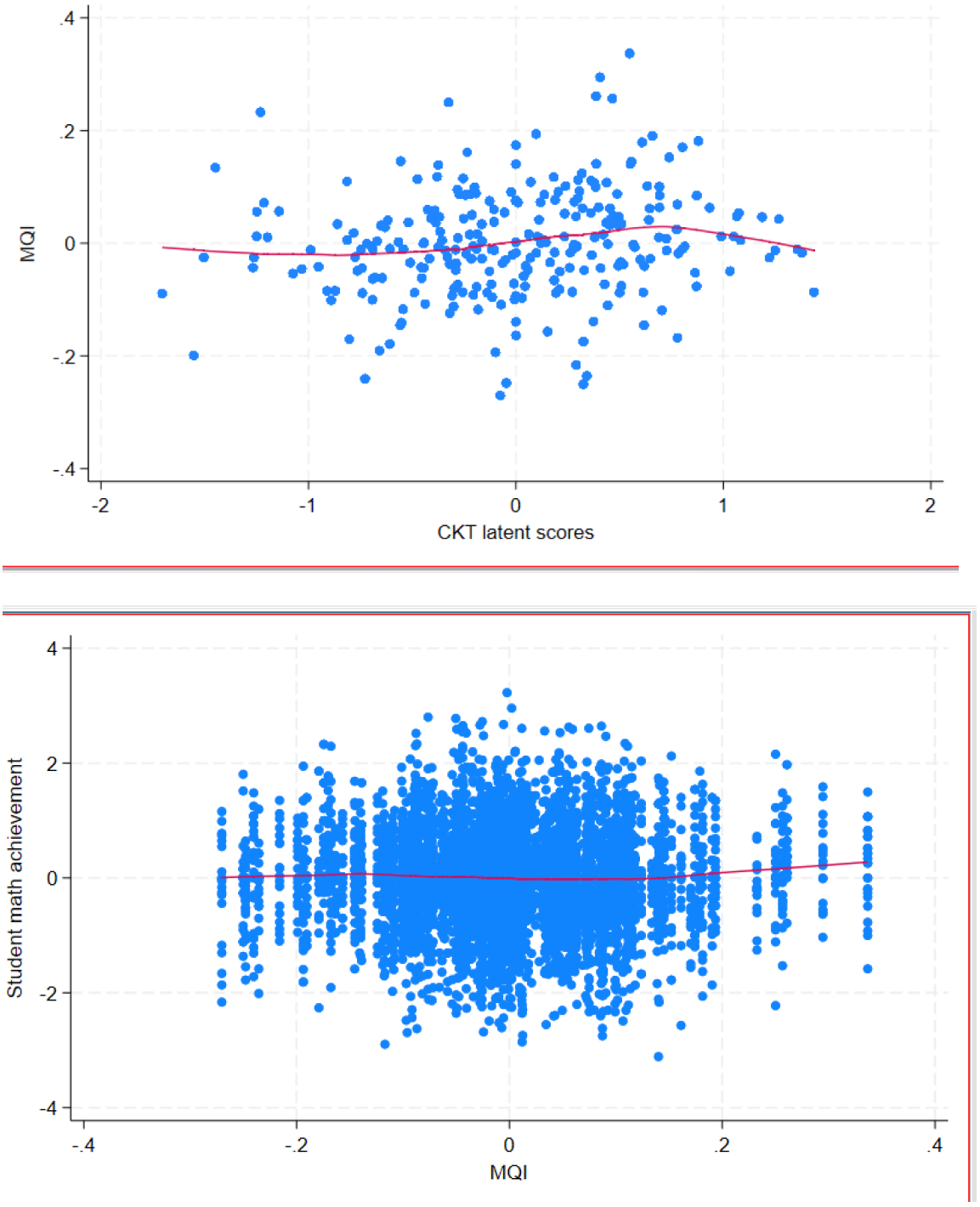


Figure 4. 3 Pairwise relationships between CKT latent scores and CLASS composite scores, and between CLASS and student achievement in elementary-school sample, all variables group mean centered

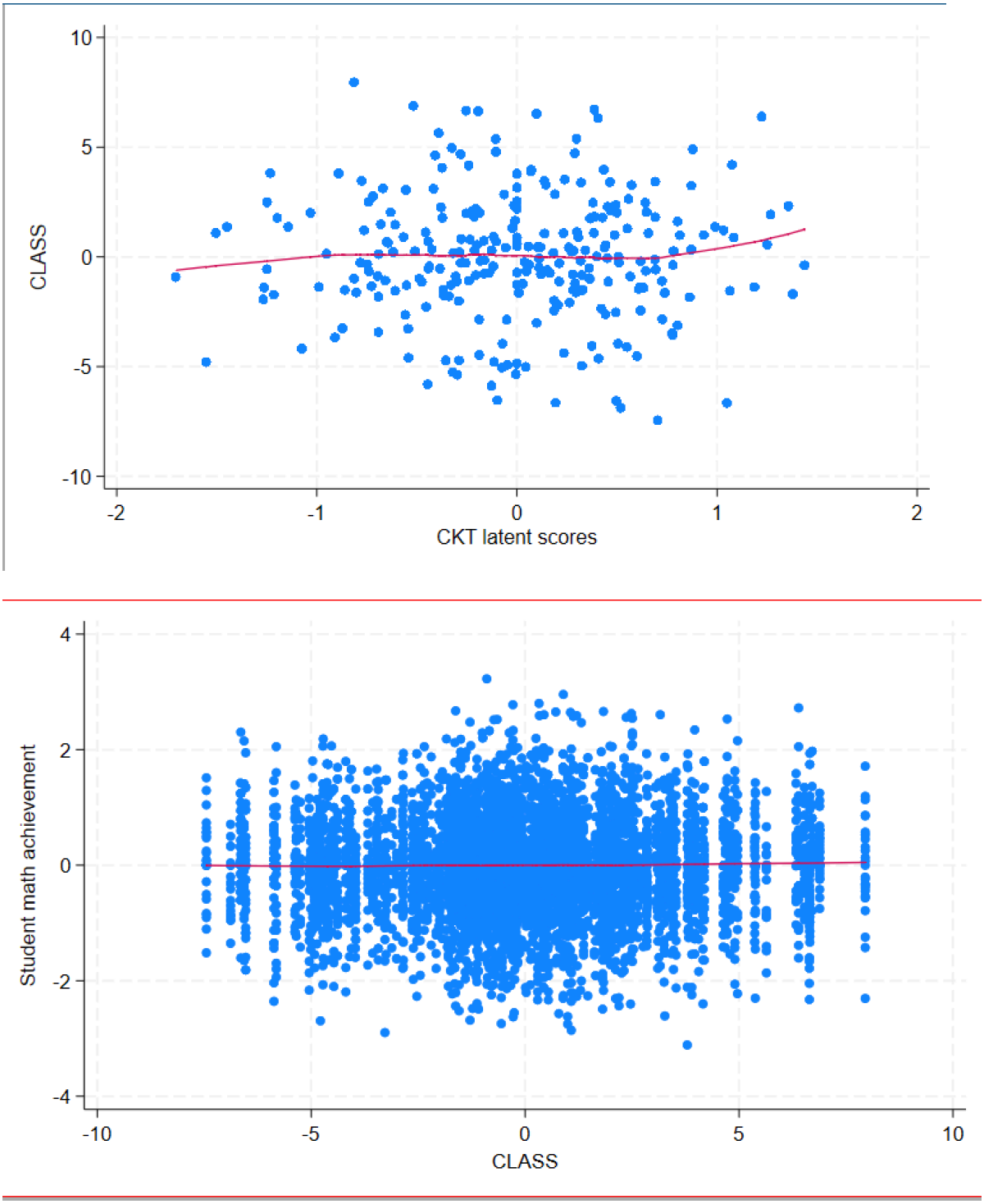


Figure 4. 4 Pairwise relationships between CKT latent scores and Tripod composite scores, and between Tripod and student achievement in elementary-school sample, all variables group mean centered

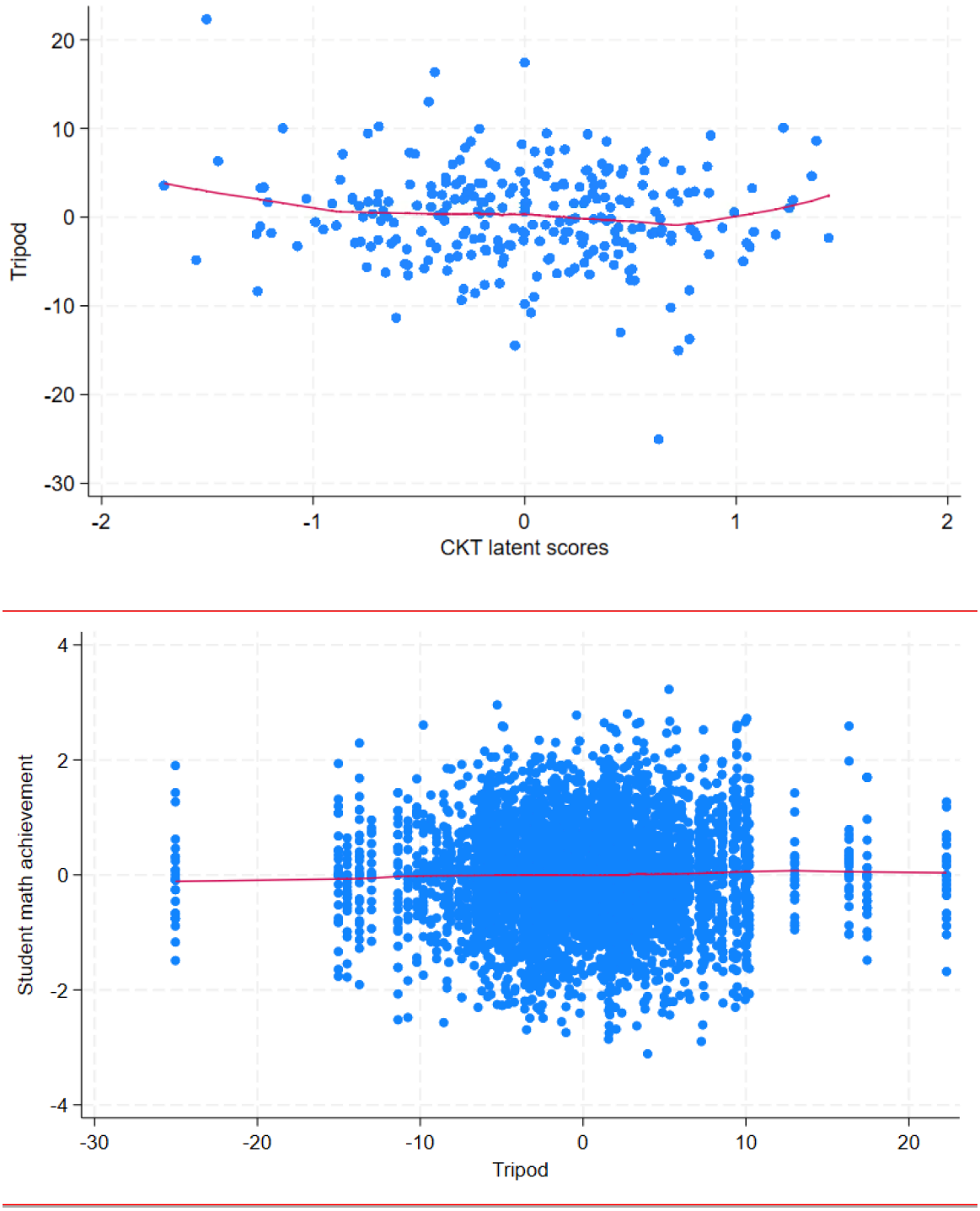


Figure 4. 5 Pairwise relationship between CKT latent scores (group mean centered) and student math achievement (group mean centered) in middle-school sample

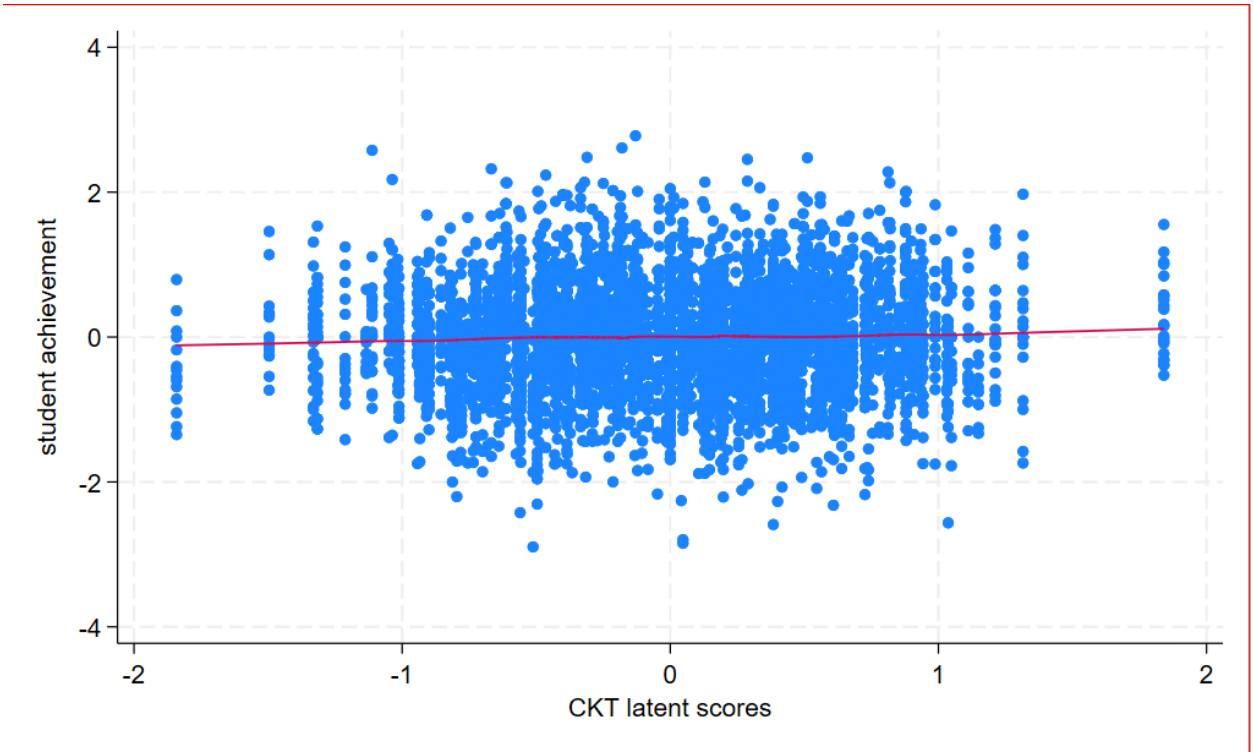


Figure 4. 6 Pairwise relationships between CKT latent scores and MQI, and between MQI and student achievement in middle-school sample, all variables group mean centered

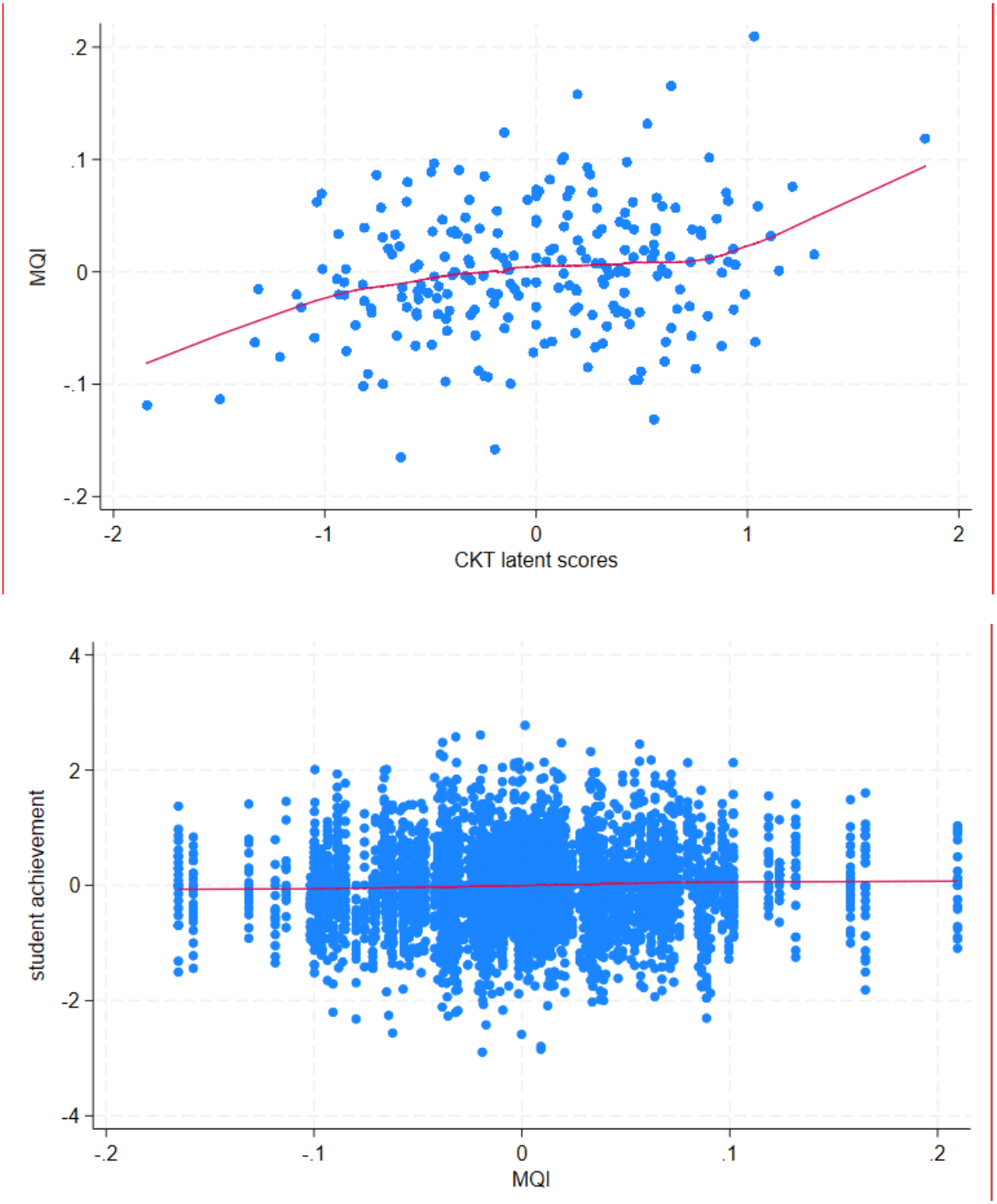
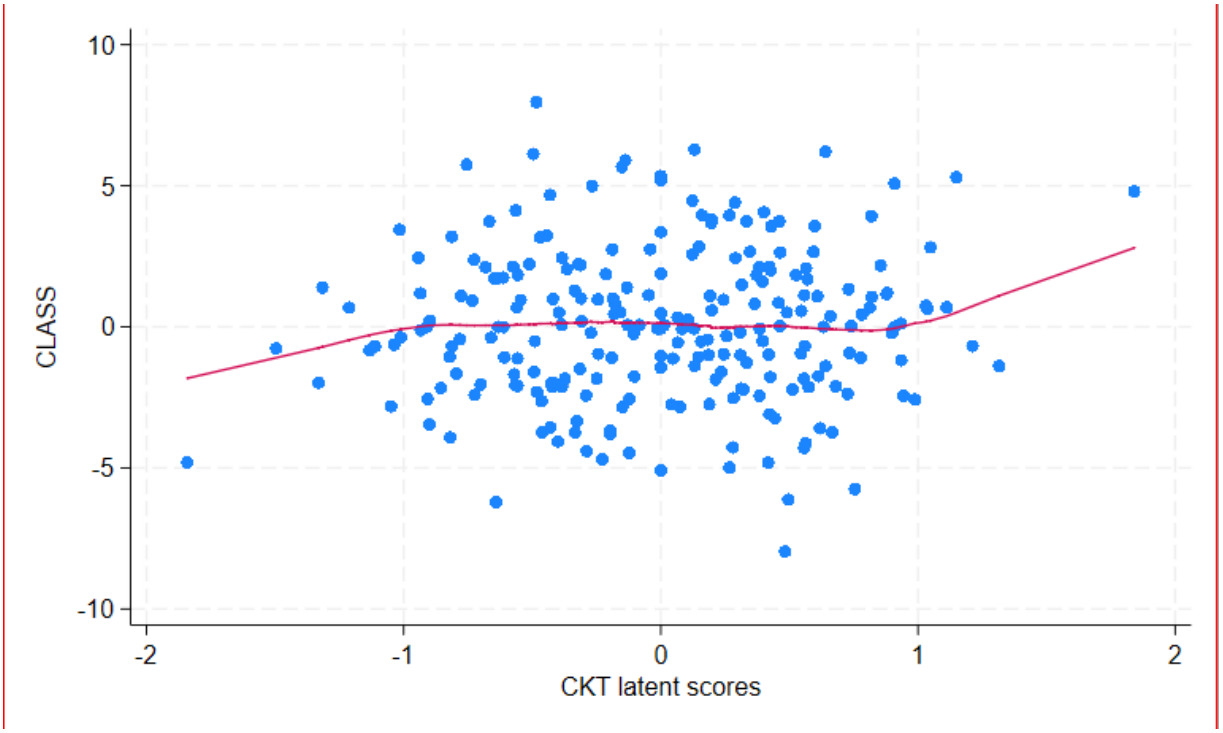


Figure 4. 7 Pairwise relationships between CKT latent scores and CLASS composite scores, and between CLASS and student achievement in middle-school sample, all variables group mean centered



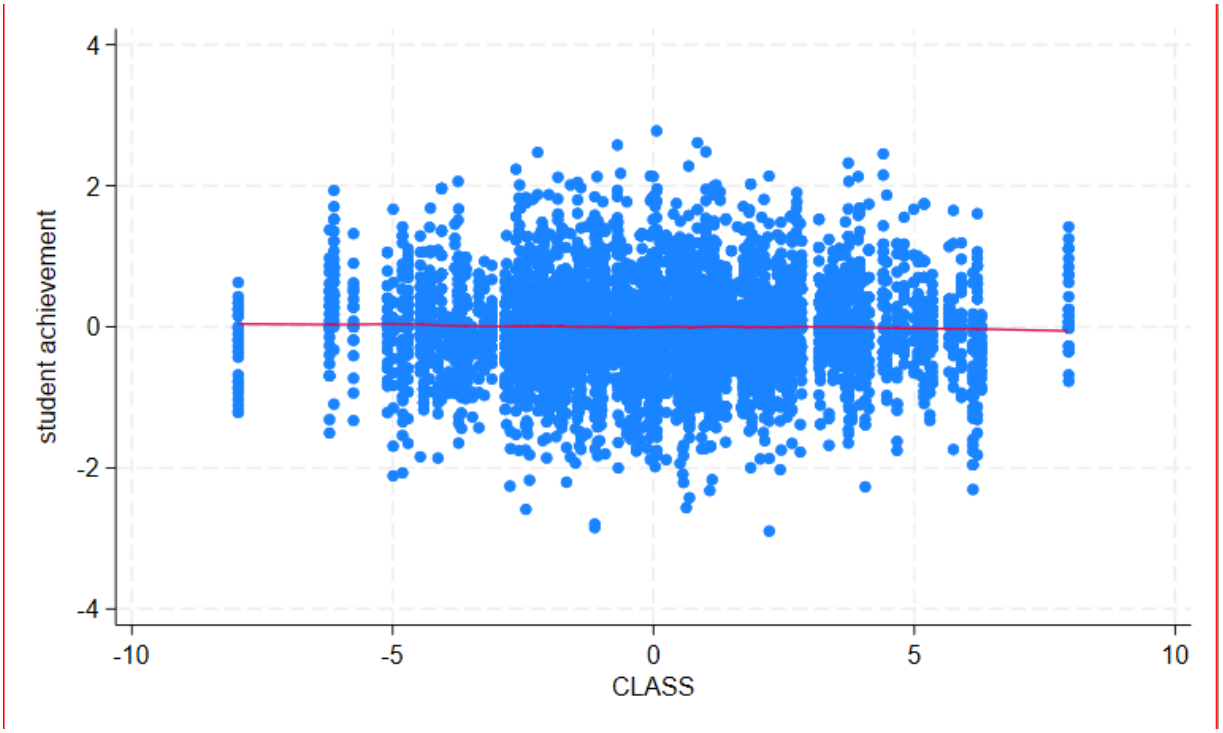
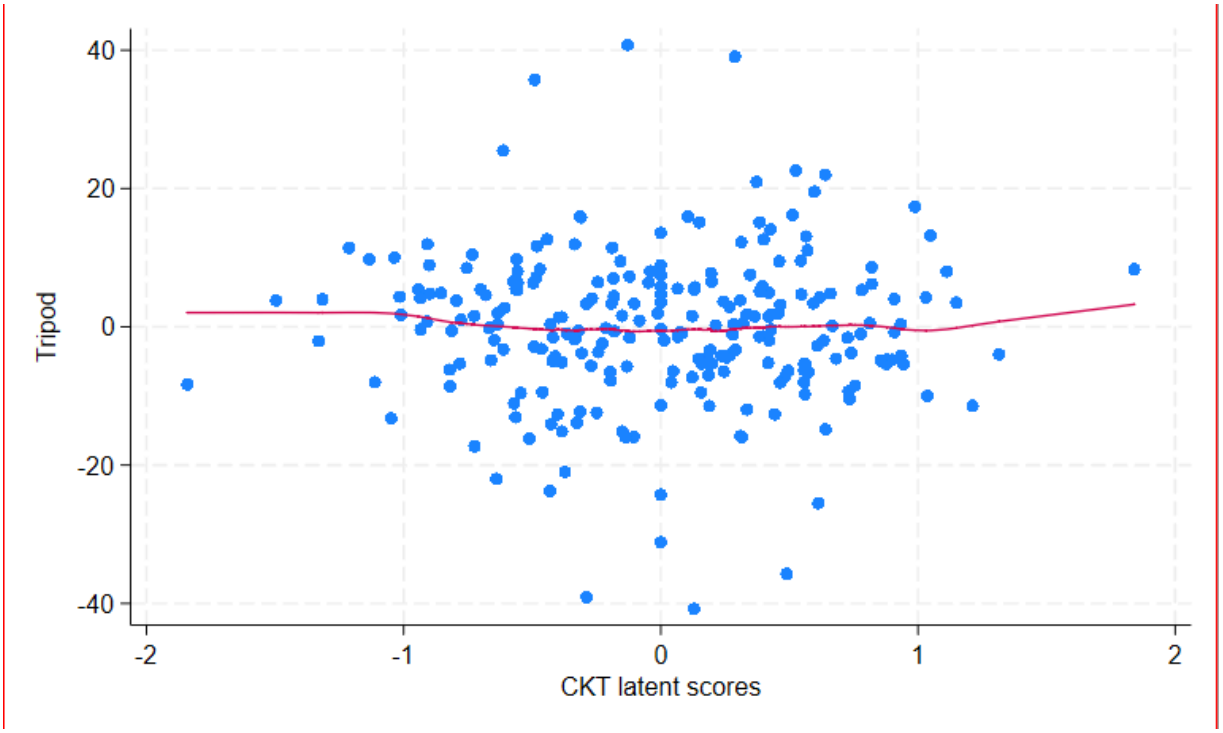


Figure 4. 8 Pairwise relationships between CKT latent scores and Tripod composite scores, and between Tripod and student achievement in middle-school sample, all variables group mean centered



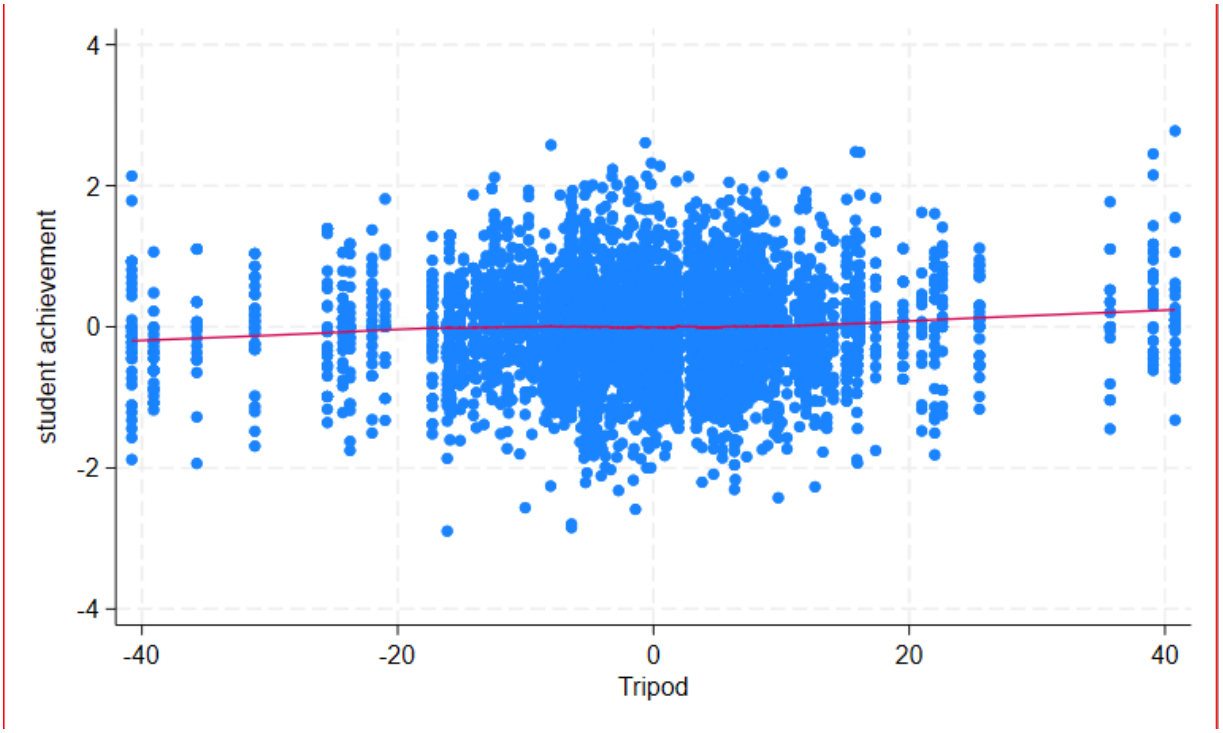


Figure 4. 9 Pairwise relationship between CKT latent scores (group mean centered) and student math achievement (group mean centered) in high-school sample

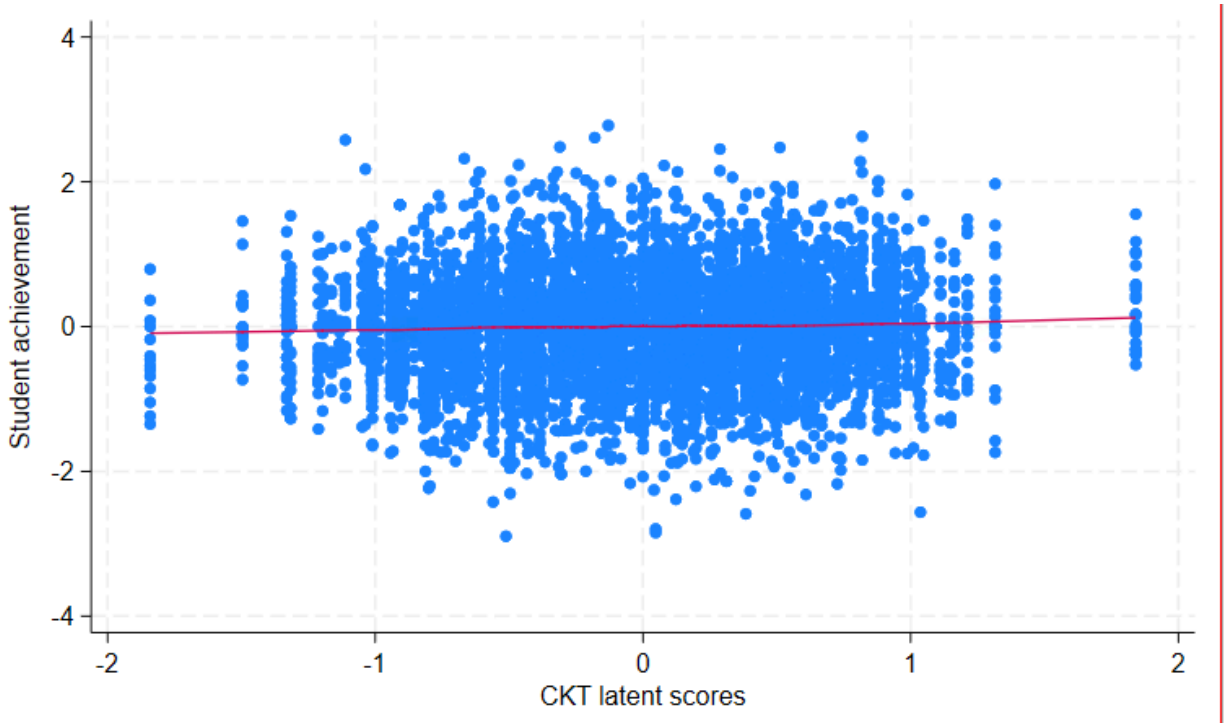


Figure 4. 10 Pairwise relationships between CKT latent scores and MQI, and between MQI and student achievement in high -school sample, all variables group mean centered

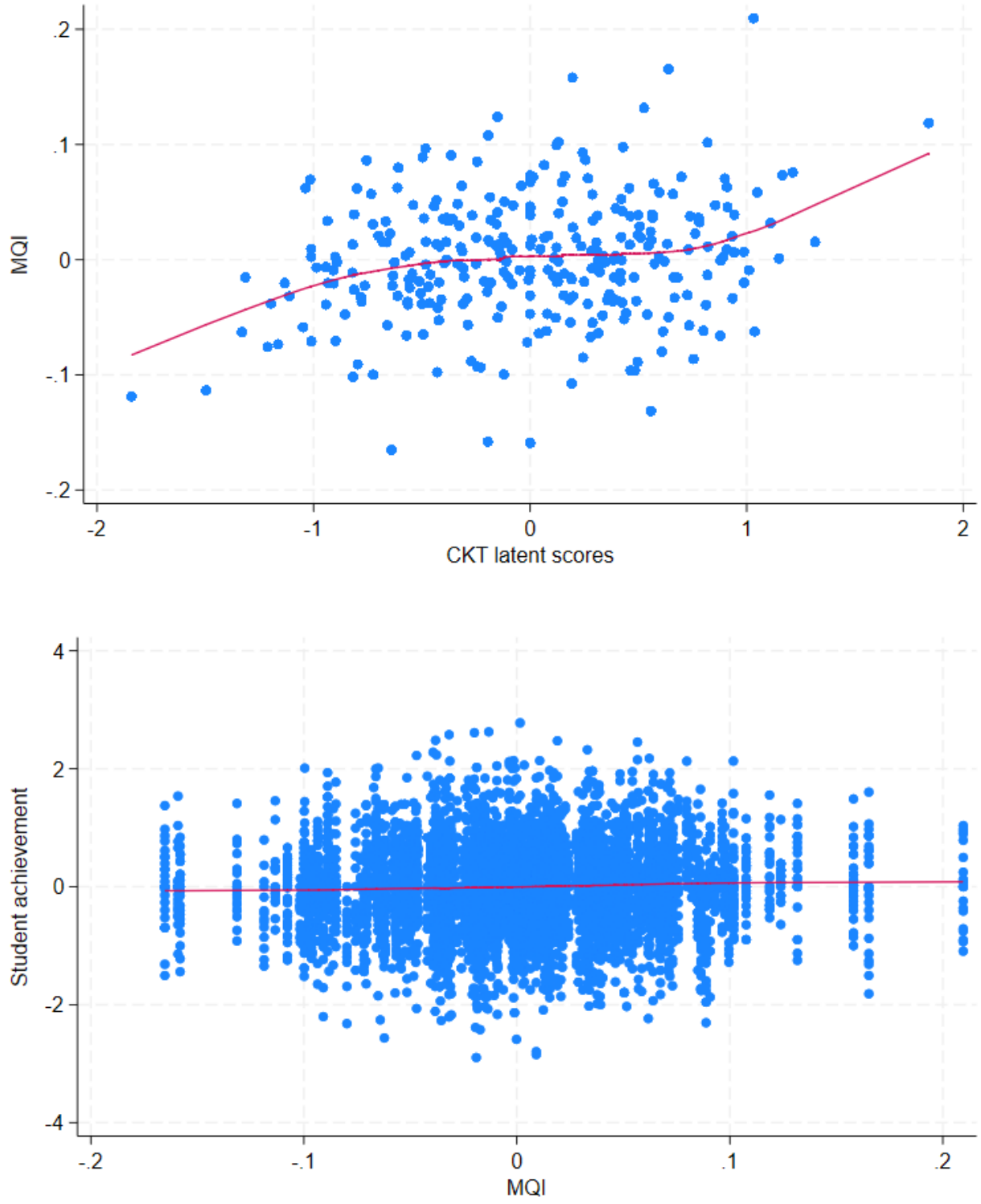


Figure 4. 11 Pairwise relationships between CKT latent scores and CLASS composite scores, and between CLASS and student achievement in high -school sample, all variables group mean

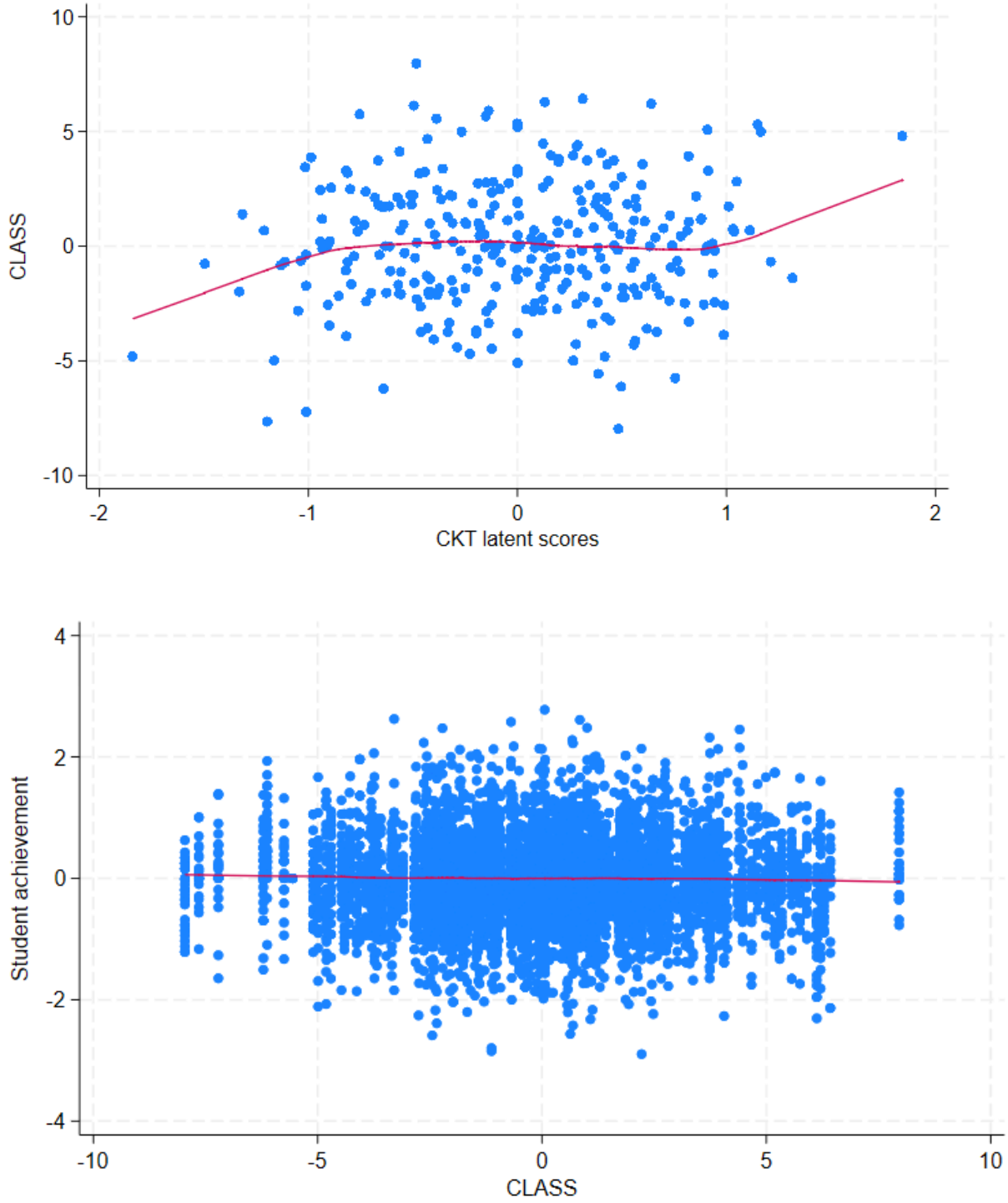
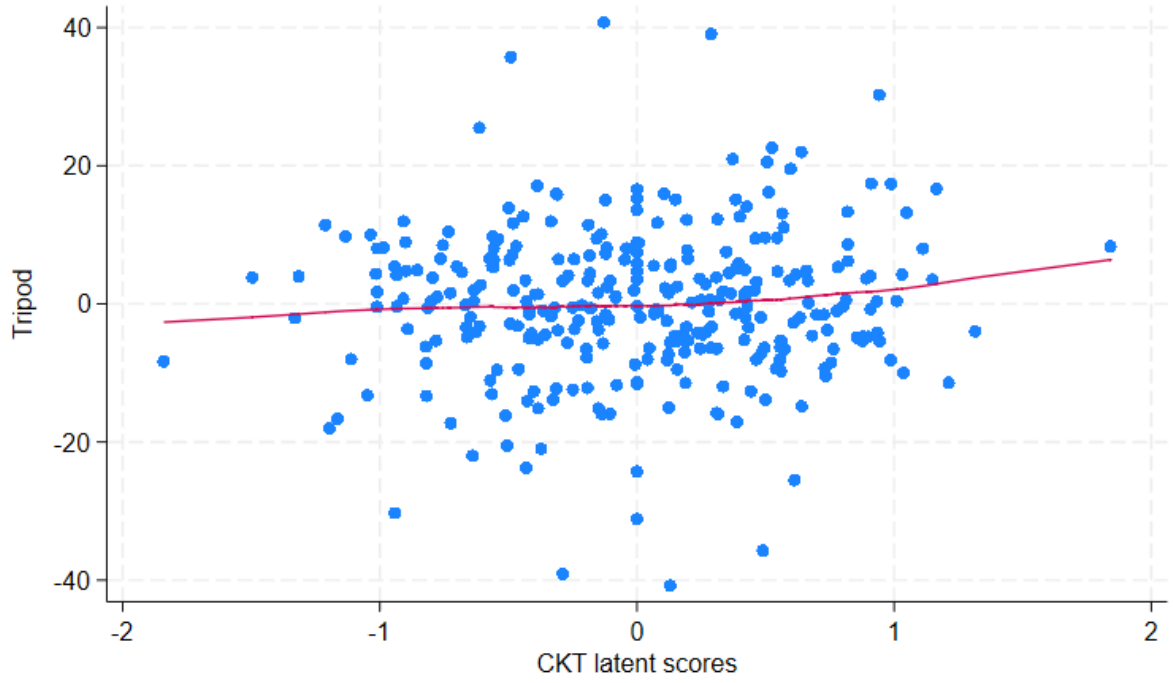
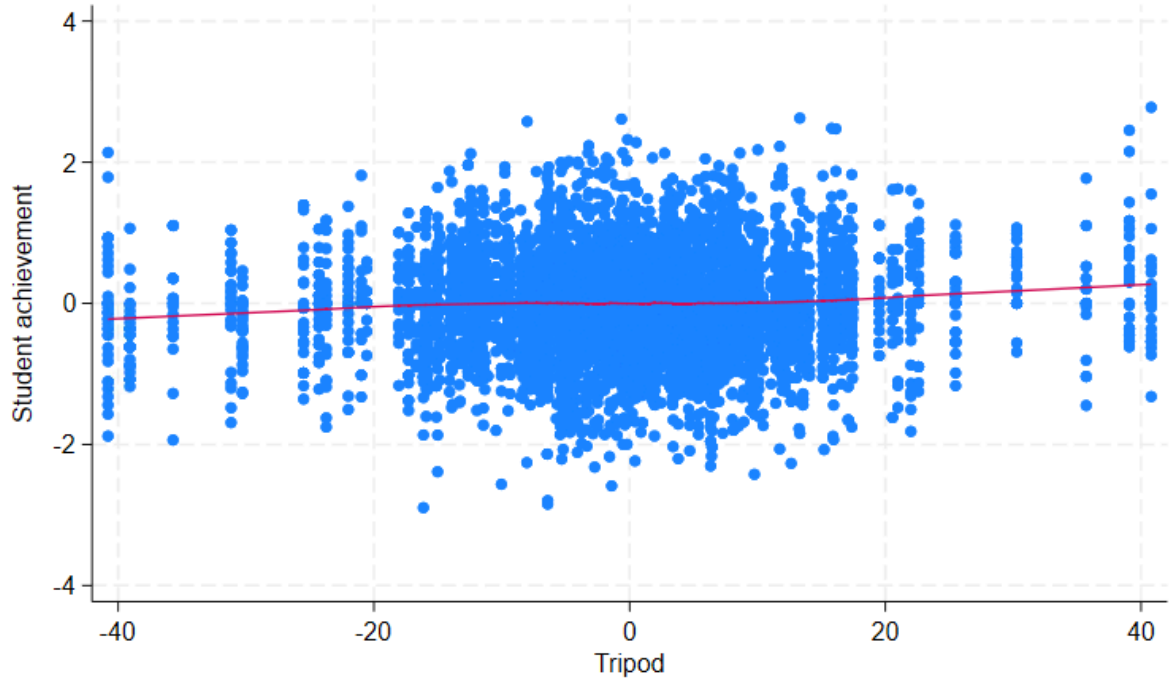


Figure 4. 12 Pairwise relationships between CKT latent scores and Tripod composite scores, and between Tripod and student achievement in high -school sample, all variables group mean centered





Reference

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2004). Schools, achievement, and inequality: A seasonal perspective. In *Summer Learning* (pp. 25–52). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781410610362-3/schools-achievement-inequality-seasonal-perspective-karl-alexander-doris-entwisle-linda-olson

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper #278. In *Economic Policy Institute*. Economic Policy Institute. https://eric.ed.gov/?id=ED516803

Ball, D. L., Hill, H. C., & Bass, H. (2005). *Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide?*

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education*, *59*(5), 389–407. https://doi.org/10.1177/0022487108324554

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Betts, J. R., Reuben, K. S., & Danenberg, A. (2000). *Equal resources, equal outcomes? The distribution of school resources and student achievement in California.* ERIC.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16–29. https://doi.org/10.1016/j.econedurev.2015.05.005

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to General and Mathematics-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments. *Educational Assessment*, *22*(2), 71–94. https://doi.org/10.1080/10627197.2017.1309274

Brophy, J., & Good, T. L. (1984). *Teacher Behavior and Student Achievement. Occasional Paper No. 73.*

Campbell, P. F., Nishio, M., Smith, T. M., Clark, L. M., Conant, D. L., Rust, A. H., DePiper, J. N., Frank, T. J., Griffin, M. J., & Choi, Y. (2014a). The Relationship Between Teachers' Mathematical Content and Pedagogical Knowledge, Teachers' Perceptions, and Student Achievement. *Journal for Research in Mathematics Education*, *45*(4), 419–459. https://doi.org/10.5951/jresematheduc.45.4.0419

Campbell, P. F., Nishio, M., Smith, T. M., Clark, L. M., Conant, D. L., Rust, A. H., DePiper, J. N., Frank, T. J., Griffin, M. J., & Choi, Y. (2014b). The relationship between teachers' mathematical content and pedagogical knowledge, teachers' perceptions, and student achievement. *Journal for Research in Mathematics Education*, *45*(4), 419–459.

Cardichon, J., Darling-Hammond, L., Yang, M., Scott, C., Shields, P. M., & Burns, D. (2020). Inequitable Opportunity to Learn: Student Access to Certified and Experienced Teachers. In *Learning Policy Institute*. Learning Policy Institute. https://eric.ed.gov/?id=ED603398

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P., & Loef, M. (1989). Using Knowledge of Children's Mathematics Thinking in Classroom Teaching: An Experimental Study. *American Educational Research Journal*, *26*(4), 499–531. https://doi.org/10.3102/00028312026004499

Charalambous, C. Y. (2010). Mathematical Knowledge for Teaching and Task Unfolding: An Exploratory Study. *The Elementary School Journal*, *110*(3), 247–278. https://doi.org/10.1086/648978

Charalambous, C. Y. (2020). Reflecting on the troubling relationship between teacher knowledge and instructional quality and making a case for using an animated teaching simulation to disentangle this relationship. *ZDM*, *52*(2), 219–240. https://doi.org/10.1007/s11858-019-01089-x

Charalambous, C. Y., Hill, H. C., Chin, M. J., & McGinn, D. (2020). Mathematical content knowledge and knowledge for teaching: Exploring their distinguishability and contribution to student learning. *Journal of Mathematics Teacher Education*, *23*(6), 579–613. https://doi.org/10.1007/s10857-019-09443-2

Charalambous, C. Y., & Litke, E. (2018). Studying instructional quality by using a content-specific lens: The case of the Mathematical Quality of Instruction framework. *ZDM*, *50*, 445–460.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2013). *MEASURING THE IMPACTS OF TEACHERS I: EVALUATING BIAS IN TEACHER VALUE-ADDED ESTIMATES*. https://irs.princeton.edu/sites/g/files/toruqf276/files/event/uploads/Impact%20of%20Teachers%20Part%201.pdf

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. *Journal of Public Economics*, *92*, 1352–1370.

Clotfelter, C., Ladd, H. F., & Vigdor, J. (2004). Teacher Quality and Minority Achievement Gaps. Working Paper Series. SAN04-04. *Terry Sanford Institute of Public Policy*.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377–392. https://doi.org/10.1016/j.econedurev.2004.06.008

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2002). *Segregation and Resegregation in North Carolina's Public School Classrooms*. https://eric.ed.gov/?id=ED471992

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *The Journal of Human Resources*, *41*(4), 778–820.

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. *Educational Researcher*, *45*(6), 378–387. https://doi.org/10.3102/0013189X16659442

Cohen, J., Hutt, E., Berlin, R., & Wiseman, E. (2022). The Change We Cannot See: Instructional Quality and Classroom Observation in the Era of Common Core. *Educational Policy*, *36*(6), 1261–1287. https://doi.org/10.1177/0895904820951114

Conger, D. (2005). Within-School Segregation in an Urban School District. *Educational Evaluation and Policy Analysis*, *27*(3), 225–244. https://doi.org/10.3102/01623737027003225

Copur-Gencturk, Y. (2015). The effects of changes in mathematical knowledge on teaching: A longitudinal study of teachers' knowledge and instruction. *Journal for Research in Mathematics Education*, *46*(3), 280–330.

Corcoran, S. P. (2007). Long-Run Trends in the Quality of Teachers: Evidence and Implications for Policy. *Education Finance and Policy*, *2*(4), 395–407.

Corcoran, S. P. (2010). Can Teachers Be Evaluated by Their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice. Education Policy for Action Series. In *Annenberg Institute for School Reform at Brown University (NJ1)*. Annenberg Institute for School Reform at Brown University. https://eric.ed.gov/?id=ED522163

Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching*. ASCD.

Danielson, C. (2008). *The Handbook for Enhancing Professional Practice: Using the Framework for Teaching in Your School*. ASCD.

Danielson, C., & Axtell, D. (2009). *Implementing the framework for teaching in enhancing professional practice*. ASCD. https://books.google.com/books?hl=en&lr=&id=AyOZCdthsQEC&oi=fnd&pg=PP11&dq=danielson+framework+for+teaching&ots=km2KhO_x3Q&sig=AeNiiHhu5-1SH7kA9l621K_P_9Y

Downey, D. B. (2023). How Does Schooling Affect Inequality in Cognitive Skills? The View From Seasonal Comparison Research. *Review of Educational Research*, 00346543231210005. https://doi.org/10.3102/00346543231210005

Downey, D. B., Von Hippel, P. T., & Broh, B. A. (2004). Are Schools the Great Equalizer? Cognitive Inequality during the Summer Months and the School Year. *American Sociological Review*, *69*(5), 613–635. https://doi.org/10.1177/000312240406900501

Entwisle, D. R., & Alexander, K. L. (1992). Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review*, 72–84.

Entwisle, D. R., & Alexander, K. L. (1994). Winter setback: The racial composition of schools and learning to read. *American Sociological Review*, 446–460.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24–28.

Figlio, D. N., & Page, M. E. (2002). School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics*, *51*(3), 497–514.

Gamoran, A., & Mare, R. D. (1989). Secondary School Tracking and Educational Inequality: Compensation, Reinforcement, or Neutrality? *American Journal of Sociology*, *94*(5), 1146–1183. https://doi.org/10.1086/229114

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., & Zhu, P. (2011). Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation. NCEE 2011-4024. *National Center for Education Evaluation and Regional Assistance*.

Gencturk, Y. C. (2012). *Teachers' mathematical knowledge for teaching, instructional practices, and student outcomes*. University of Illinois at Urbana-Champaign.

https://search.proquest.com/openview/210f627c02cf344c40daee56c95d881c/1?pq-origsite=gscholar&cbl=18750
Gill, B., Shoji, M., Coen, T., & Place, K. (2016). The content, predictive power, and potential bias in five widely used teacher observation instruments. *Mathematica Policy Research*. https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2017191.pdf
Gitomer, D. H. (2019). Evaluating instructional quality. *School Effectiveness and School Improvement*, *30*(1), 68–78. https://doi.org/10.1080/09243453.2018.1539016
Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven Playing Field? Assessing the Teacher Quality Gap Between Advantaged and Disadvantaged Students. *Educational Researcher*, *44*(5), 293–307. https://doi.org/10.3102/0013189X15592622
Hamre, B. K., & Pianta, R. C. (2007). *Learning opportunities in preschool and early elementary classrooms.*
Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagami, A. (2013). Teaching through Interactions: Testing a Developmental Framework of Teacher Effectiveness in over 4,000 Classrooms. *The Elementary School Journal*, *113*(4), 461–487. https://doi.org/10.1086/669616
Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 US early childhood and elementary classrooms. *Foundation for Childhood Development*, *30*(2008).
Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1999). *Do higher salaries buy better teachers?* National bureau of economic research.
Hanushek, E. A., Kain, J., O'Brien, D., & Rivkin, S. G. (2005). *The market for teacher quality*. National Bureau of Economic Research Cambridge, Mass., USA. https://www.nber.org/papers/w11154
Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*, *100*(2), 267–271. https://doi.org/10.1257/aer.100.2.267
Hanushek, E. A., & Rivkin, S. G. (2012). The Distribution of Teacher Quality and Implications for Policy. *Annu. Rev. Econ*, *4*, 131–157.
Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. John Wiley & Sons.
Hill, H. C. (2007). Mathematical knowledge of middle school teachers: Implications for the No Child Left Behind policy initiative. *Educational Evaluation and Policy Analysis*, *29*(2), 95–114.
Hill, H. C. (2010). The nature and predictors of elementary teachers' mathematical knowledge for teaching. *Journal for Research in Mathematics Education*, 513–545.
Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, *26*(4), 430–511. https://doi.org/10.1080/07370000802177235
Hill, H. C., & Chin, M. (2018). Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Educational Research Journal*, *55*(5), 1076–1112.
Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831.

Hill, H. C., Litke, E., & Lynch, K. (2018). Learning Lessons from Instruction: Descriptive Results from an Observational Study of Urban Elementary Classrooms. *Teachers College Record: The Voice of Scholarship in Education*, *120*(12), 1–46. https://doi.org/10.1177/016146811812001207

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, *105*(1), 11–30.

Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher Quality and Quality Teaching: Examining the Relationship of a Teacher Assessment to Practice. *American Journal of Education*, *118*(4), 489–519. https://doi.org/10.1086/666380

Hill, H., & Grossman, P. (2013). Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems. *Harvard Educational Review*, *83*(2), 371–384. https://doi.org/10.17763/haer.83.2.d11511403715u376

Hong, G. (2015). *Causality in a social world: Moderation, mediation and spill-over*. John Wiley & Sons.

Huber, M., Hsu, Y.-C., Lee, Y.-Y., & Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, *35*(7), 814–840.

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Jacob, R., Hill, H., & Corey, D. (2017). The impact of a professional development program on teachers' mathematical knowledge for teaching, instruction, and student achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 379–407.

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education*, *38*(3), 258–288.

Kalogrides, D., & Loeb, S. (2013). Different Teachers, Different Peers: The Magnitude of Student Sorting Within Schools. *Educational Researcher*, *42*(6), 304–316. https://doi.org/10.3102/0013189X13495087

Kalogrides, D., Loeb, S., & Beteille, T. (2011). Power Play? Teacher Characteristics and Class Assignments. Working Paper 59. In *National Center for Analysis of Longitudinal Data in Education Research*. National Center for Analysis of Longitudinal Data in Education Research. https://eric.ed.gov/?id=ED519993

Kalogrides, D., Loeb, S., & Béteille, T. (2013). Systematic Sorting: Teacher Characteristics and Class Assignments. *Sociology of Education*, *86*(2), 103–123. https://doi.org/10.1177/0038040712456555

Kane, T. (2012). Capturing the dimensions of effective teaching: Student achievement gains, student surveys, and classroom observations. *Education Next*, *12*(4), 34–42.

Kane, T. J., & Staiger, D. O. (2012a). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kane, T. J., & Staiger, D. O. (2012b). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587–613.

Kelcey, B. (2011). Assessing the Effects of Teachers' Reading Knowledge on Students' Achievement Using Multilevel Propensity Score Stratification. *Educational Evaluation and Policy Analysis*, *33*(4), 458–482. https://doi.org/10.3102/0162373711415262

Kelcey, B., Hill, H. C., & Chin, M. J. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: A multilevel quantile mediation analysis. *School Effectiveness and School Improvement*, *30*(4), 398–431.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568–589.

Knight, D. S. (2019). Are school districts allocating resources equitably? The Every Student Succeeds Act, teacher experience gaps, and equitable resource allocation. *Educational Policy*, *33*(4), 615–649.

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod student survey. *Educational Assessment*, *22*(4), 253–274.

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*(1), 25–47. https://doi.org/10.1007/s10857-010-9140-1

Liu, J., & Cohen, J. (2021). Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods. *Educational Evaluation and Policy Analysis*, *43*(4), 587–614. https://doi.org/10.3102/01623737211009267

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics*, *9*(3). https://doi.org/10.1214/15-AOAS833

Loeb, S. (2000). How teachers' choices affect what a dollar can buy: Wages and quality in K-12 schooling. *Proceedings from the Symposium on the Teaching Workforce. Albany, New York, Education Finance Research Consortium, November*, *8*.

Lynch, K., Chin, M., & Blazar, D. (2017). Relationships between Observations of Elementary Mathematics Instruction and Student Achievement: Exploring Variability across Districts. *American Journal of Education*, *123*(4), 615–646. https://doi.org/10.1086/692662

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation Analysis. *Annual Review of Psychology*, *58*, 593. https://doi.org/10.1146/annurev.psych.58.110405.085542

Metzler, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics*, *99*(2), 486–496. https://doi.org/10.1016/j.jdeveco.2012.06.002

Mickelson, R. A., & Everett, B. J. (2008). Neotracking in North Carolina: How High School Courses of Study Reproduce Race and Class-Based Stratification. *Teachers College Record: The Voice of Scholarship in Education*, *110*(3), 535–570. https://doi.org/10.1177/016146810811000306

Milanowski, A. (2004). The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33–53.

Morgan, P. R., & McPartland, J. M. (1981). *The Extent of Classroom Segregation within Desegregated Schools*. https://eric.ed.gov/?id=ED210405

Mu, J., Bayrak, A., & Ufer, S. (2022). Conceptualizing and measuring instructional quality in mathematics education: A systematic literature review. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.994739

Perry, R., & Lewis, C. (2011). *Improving the Mathematical Content Base of Lesson Study Summary of Resultsi*.

Phillips, S. F., Ferguson, R. F., & Rowley, J. F. S. (2021). Do They See What I See? Toward a Better Understanding of the 7Cs Framework of Teaching Effectiveness. *Educational Assessment*, 1–19. https://doi.org/10.1080/10627197.2020.1858784

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher*, *38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Pianta, R. C., Hamre, B. K., & Downer, J. (2011). *Aligning measures of quality with professional development goals and goals for children's development.* https://psycnet.apa.org/record/2011-10988-013

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, *121*(2), 183–212.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Raudenbush, S. W., & Jean, M. (2015). To What Extent Do Student Perceptions of Classroom Quality Predict Teacher Value Added. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems* (1st ed., pp. 170–202). Wiley. https://doi.org/10.1002/9781119210856.ch6

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*, *6*(1), 43–74. https://doi.org/10.1162/EDFP_a_00022

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., & Gallagher, L. P. (2010). Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-Scale Studies. *American Educational Research Journal*, *47*(4), 833–878. https://doi.org/10.3102/0002831210367426

Rowley, J. F. S., Phillips, S. F., & Ferguson, R. F. (2019). The stability of student ratings of teacher instructional practice: Examining the one-year stability of the 7Cs composite. *School Effectiveness and School Improvement*, *30*(4), 549–562.

Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2010). Problem Implementation as a Lever for Change: An Experimental Study of the Effects of a Professional Development Program on Students' Mathematics Learning. *Journal of Research on Educational Effectiveness*, *4*(1), 1–24. https://doi.org/10.1080/19345747.2010.498562

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, *7*, 110.

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, *48*(1), 29–40. https://doi.org/10.1007/s11858-016-0765-0

Senden, B., Nilsen, T., & Blömeke, S. (2022). 5. Instructional Quality: A Review of Conceptualizations, Measurement Approaches, and Research Findings. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of Analyzing Teaching Quality* (pp. 140–172). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021-05

Shechtman, N., Roschelle, J., Haertel, G., & Knudsen, J. (2010a). Investigating links from teacher knowledge, to classroom practice, to student learning in the instructional system of the middle-school mathematics classroom. *Cognition and Instruction*, *28*(3), 317–359.

Shechtman, N., Roschelle, J., Haertel, G., & Knudsen, J. (2010b). Investigating links from teacher knowledge, to classroom practice, to student learning in the instructional system of the middle-school mathematics classroom. *Cognition and Instruction*, *28*(3), 317–359.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis*, *38*(2), 199–221.

Standing, K., & Lewis, L. (2021). Pre-COVID Ability Grouping in US Public School Classrooms. Data Point. NCES 2021-139. *National Center for Education Statistics*.

Steele, J. L., Murnane, R. J., & Willett, J. B. (2010). Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California. *Journal of Policy Analysis and Management*, *29*(3), 451–478.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research*, *86*(4), 849–899.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review*, *100*(2), 256–260.

Vanfossen, B. E., Jones, J. D., & Spade, J. Z. (1987). Curriculum tracking and status maintenance. *Sociology of Education*, 104–122.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What Can Student Perception Surveys Tell Us About Teaching? Empirically Testing the Underlying Structure of the Tripod Student Perception Survey. *American Educational Research Journal*, *53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Appendices

Alternative Model for Step 3. Analyzing the mediated effect on the outcome via the mediator and the treatment-by-mediator interaction conditioning on the covariates

Level-1 Model (Student)

$$Y_{ijrk} = \varphi_{0jrk} + \varphi_{1jrk}\,\text{pre-test}_{ijrk} + \varphi_{2jrk} X_{ijrk} + e_{ijrk}$$

$$e_{ijrk} \sim N(0, \sigma^2_e)$$

Level-2 Model (Classroom)

$$\varphi_{0jrk} = \pi_{00rk} + \pi_{01rk}\,\text{InsQ}_{jrk} + \pi_{02rk}\,\text{CKT}_{jrk} + v_{0jrk}$$

$$\varphi_{1jrk} = \pi_{10rk}$$

$$\varphi_{2jrk} = \pi_{20rk}$$

$$v_{0jrk} \sim N(0, \sigma^2_v)$$

Level-3 Model (Randomization block)

$$\pi_{00rk} = \theta_{000k} + \theta_{001k}\text{Grade}_{rk} + \varepsilon_{00rk}$$

$$\pi_{01rk} = \theta_{010k}$$

$$\pi_{02rk} = \theta^3_{020k}$$

$$\pi_{10rk} = \theta^3_{100k}$$

$$\pi_{20rk} = \theta^3_{200k}$$

$$\varepsilon_{00rk} \sim N(0, \sigma^2_\varepsilon)$$

Level-4 Model (School)

$$\theta_{000k} = \beta_{0000} + \beta_{0100}\,\text{District}_k + d_{000k}$$

$$\theta_{001k} = \beta_{0010}$$

$$\theta_{010k} = \beta_{0100}$$

$$\theta_{020k} = \beta_{0200}$$

$$\theta_{100k} = \beta_{1000}$$

$$\theta_{200k} = \beta_{2000}$$

$$d_{00rk} \sim N(0, \sigma^2_d)$$

Mixed Model:
$$Y_{ijrk} = \beta_{0000} + \beta_{0100} CKT_{jrk} + \beta_{0200} InsQ_{jrk} + \beta_{0300} InsQ_{jrk} CKT_{jrk}$$
$$+ \beta_{1000} \text{pre-test}_{ijrk} + \beta_{2000} X_{ijrk}$$
$$+ \beta_{0100} District_k + \beta_{0010} Grade_{rk}$$
$$+ d_{000k} + \varepsilon_{00rk} + v_{0jrk} + e_{ijrk}$$

Note: Subscripts *ijrk* indicate student *i* in classroom *j* of matched cluster *m* of school *k*; X_{ijrk} indicates a vector of student-level characteristics, including their prior test scores (math scores in AY 2009-2010), eligibility of F/R lunch, racial identities, ELL status, and Special Education status. $Grade_{rk}$ represents a vector of indicator variables for grade level, e.g., in middle school sample, $Grade_{ij} = [Grade5_{rk}, Grade6_{rk}]^T$ where $Grade_{5rk}$ and $Grade_{6rk}$ are binary variables, indicating the classroom j in randomization block r is at 5th grade level. $District_k$ represents a vector of indicator variables for districts, $District_{0j} = [District2_k, \ldots, District5_k]^T$, where $District2_k$ to $District5_k$ are binary variables. Residual variance components d, ε, v, e represent variation at the school-, randomization block-, classroom-, and student-levels and are assumed to follow zero mean normal distributions with variance σ^2_d, σ^2_ε, σ^2_v, σ^2_e respectively.

Appendix Tables

Table A2. 1 Model Specifications for Weighted Regression

Unadjusted model	Adjusted model
Level-1 Model (Teacher) $CKT_{ij} = \pi_{0j} + \pi_{1j}Grade_{ij} + e_{ij}$ $e_{ij} \sim N(0, \sigma_e^2)$	Level-1 Model (Teacher) $CKT_{jk} = \pi_{0k} + \pi_{1k}X_{jk`} + \pi_{2k}Grade_{jk} + e_{jk}$ $e_{jk} \sim N(0, \sigma_e^2)$
Level-2 Model (School) $\pi_{0j} = \beta_{00} + \beta_{01}\ District_k + r_{0j}$ $\pi_{1j} = \beta_{10}$ $r_{0j} \sim N(0, \sigma_r^2)$	Level-2 Model (School) $\pi_{0k} = \beta_{00} + \beta_{01}\ X_k + \beta_{02}\ District_k + r_{0k}$ $\pi_{1k} = \beta_{10}$ $\pi_{2k} = \beta_{20}$ $r_{0k} \sim N(0, \sigma_r^2)$
Mixed Model $CKT_{ij} = \beta_{00} + \beta_{01}\ District_k + \beta_{10}Grade_{jk} + r_{0j} + e_{jk}$	Mixed Model $CKT_{jk} = \beta_{00} + \beta_{01}X_k + \beta_{10}X_{jk} + \beta_{02}District_k + \beta_{20}\ Grade_{jk} + r_{0k} + e_{jk}$

Note: Here, subscripts i, j represent classroom i in school j. Coefficients π, β represents classroom-level coefficients, and school-level coefficients respectively. $Grade_{ij}$ represents a vector of indicator variables for grade level, $Grade_{ij} = [Grade5_{ij}, \ldots, Grade9_{ij}]^T$ where $Grade4_{ij}$ to $Grade8_{ij}$ are dummy variables. Any one of the grade-level dummy indicates the teacher i in school j teaches the 5th to the 9th grades correspondingly; all equal to 0 indicates the teacher teaches the 4th grade. π_{1j} is a vector of coefficients for vector $Grade_{ij}$. Similarly, $District_{0j}$ represents a vector of indicator variables for district, $District_{0j} = [District2_{0j}, \ldots, District6_{0j}]^T$ where $District2_{0j}$ to $District6_{0j}$ are dummy variables. Any one of district dummy variables equal to 1 indicates school j being in District 2 to 6 correspondingly; all equal to 0 indicates school j being in District 1. β_{01} is a vector of coefficients for vector $District_{0j}$. Residual variance components e, r represent variation at the classroom-, and school-levels and are assumed to follow zero mean normal distributions with variance σ_e^2, σ_r^2 respectively.

Table A2. 2 Natural variation by grade, by state and variance decomposition of full sample and subsamples of various levels of schools (Weighted regression)

VARIABLES	(1) Full sample	(2) Elementary school (4th to 5th)	(3) Middle school subsample (6th to 8th)	(4) High school subsample (9th)
Intercept	-.001	.177	.065	.173
	(.008)	(.017)	(.007)	(.014)
Grade indicators				
4th grade	baseline	baseline	-	-
5th grade	.060	.052	-	-
	(.005)	(.005)		
6th grade	.091	-	baseline	-
	(.011)			
7th grade	.068	-	-.030	-
	(.012)		(.015)	
8th grade	.448***	-	.342**	-
	(.011)		(.012)	
9th grade	.546***	-	-	-
	(.010)			
District indicators	YES	YES	YES	YES
District 1	baseline	baseline	baseline	baseline
District 2	.296*	.094	-	.691*
	(.047)	(.111)		(.088)
District 3	-.691***	-.959***	-.747***	.159
	(.009)	(.020)	(.017)	(.036)
District 4	.093	-.144	.208	.557*
	(.010)	(.021)	(.028)	(.059)
District 5	-.134	-.328*	-.079	.201
	(.009)	(.023)	(.024)	(.031)
District 6	-.415**	-	-.381*	-
	(.021)		(.021)	
Within-school variation of the random intercept	.644	.612	.687	.609
Between-school variation of the random intercept	.070	.030	.077	.073
Intraclass correlation	.108	.046	.101	.108
Observations				
Number of classrooms	908	394	373	141
Number of schools	267	109	100	68

Table A2. 3 Inequality in CKT distribution by prior performance levels (Weighted regression)

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Average Math scores in 2009 of students taught by the teacher	-.098	.086	-.084
	(.031)	(.012)	(.113)
School average Math scores in 2009	.276*	.638***	.681**
	(.020)	(.015)	(.047)
Intercept	.142	-.013	.226
	(.016)	(.009)	(.015)
Within-school variation of the random intercept	.606	.698	.612
Between-school variation of the random intercept	.027	.008	.000
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Number of classrooms	394	373	131
Number of schools	109	100	67

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Table A2. 4 Inequality in CKT distribution by Free/Reduced-priced Lunch Status (Weighted regression)

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of F/R eligible students taught by the teacher	-.799	-.990	2.183*
	(.784)	(.242)	(1.158)
Proportion of F/R eligible students in school	-.445	-.394	-1.710***
	(.072)	(.085)	(.189)
Intercept	.223*	.032	.238
	(.015)	(.008)	(.018)
Within-school variation of the random intercept	.564	.704	.538
Between-school variation of the random intercept	.049	.062	.036
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Number of classrooms	290	315	123
Number of schools	85	87	60

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Note: The sample size for regressions including F/R lunch statuses is notably smaller compared with other regressions. This vast difference came from the missing information of students' F/R lunch statuses in the administrative records of an entire school district.

Table A2. 5 Inequality in CKT distribution by Minority Status (Weighted regression)

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of minority students taught by the teacher	-.208	-.625	-.134
	(.535)	(.289)	(.842)
Proportion of minority students in School	-.366	-.903***	-1.511**
	(.045)	(.071)	(.278)
Intercept	.102	-.097	-.139
	(.018)	(.011)	(.031)
Within-school variation of the random intercept	.610	.692	.605
Between-school variation of the random intercept	.026	.036	.027
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Number of classrooms	394	373	141
Number of schools	109	100	68

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Table A2. 6 Inequality in CKT distribution by English Language Learner Status (Weighted regression)

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of ELL students taught by the teacher	.678	-.656	-.464
	(.153)	(.162)	(1.150)
Proportion of ELL students in School	-.560	-.122	-2.619**
	(.212)	(.654)	(.302)
Intercept	.179	.060	.109
	(.015)	(.009)	(.014)
Within-school variation of the random intercept	.607	.681	.621
Between-school variation of the random intercept	.027	.078	.017
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Number of classrooms	394	373	141
Number of schools	109	100	68

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Table A2. 7 Inequality in CKT distribution by Special Education Status (Weighted regression)

VARIABLES	(1) Elementary schools	(2) Middle Schools	(3) High Schools
Proportion of Special Ed students taught by the teacher	.608	-.192	-.592
	(.552)	(.170)	(.830)
Proportion of Special Ed students in School	.927	-1.841	-6.581**
	(1.039)	(2.091)	(5.929)
Intercept	.142	.026	.264
	(.017)	(.009)	(.013)
Within-school variation of the random intercept	.612	.693	.598
Between-school variation of the random intercept	.028	.063	.047
Grade indicators	YES	YES	YES
District indicators	YES	YES	YES
Observations			
Number of classrooms	393	373	141
Number of schools	109	100	68

Standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05

Table A3. 1 CLASS theoretical domains

Theoretical domains	**Items**
EMOTIONAL SUPPORT	Positive Climate
	Negative Climate
	Teacher Sensitivity
	Regards for Student Perspective
CLASSROOM ORGANIZATION	Behavior Management
	Productivity
	Instructional Learning Formats
INSTRUCTIONAL SUPPORT	Content Understanding
	Analysis and Problem Solving
	Quality of Feedback
	Instructional Dialogue
STUDENT ENGAGEMENT	Student Engagement

Table A3. 2 MQI domains

MQI Items
Classroom Work Connected to Mathematics
Errors and Imprecision*(Not in Elementary School Form)
Explicitness and Thoroughness
Richness of Mathematics
Student Participation in Meaning Making and Reasoning
Working with Students and Mathematics

Table A3. 3 Tripod items

Theoretical domains	Elementary survey code	Secondary survey code	Item description
CARE	126	a10	My teacher in this class makes me feel that s/he really cares about me.
	180	b146	My teacher seems to know if something is bothering me.
		b34	My teacher really tries to understand how students feel about things.
CONFER	168	b129	My teacher wants us to share our thoughts.
	159	b154	My teacher gives us time to explain our ideas.
	176	b155	Students speak up and share their ideas about class work.
CAPTIVATE		b29	My teacher makes learning enjoyable.
		b44	My teacher makes lessons interesting.
	108	b89	I like the ways we learn in this class.
		b141	This class does not keep my attention – I get bored.
CLARIFY	96	b1	If you don't understand something, my teacher explains it another way.
	185	b130	My teacher knows when the class understands, and when we do not.
		b136	When s/he is teaching us, my teacher thinks we understand even when we don't. 157
	97	b17	My teacher has several good ways to explain each topic that we cover in this class.
	98	b80	My teacher explains difficult things clearly.
	111	b90	In this class, we learn to correct our mistakes.
	165	b147	My teacher checks to make sure we understand what s/he is teaching us.
		b58	We get helpful comments to let us know what we did wrong on assignments.
		b83	The comments that I get on my work in this class help me understand how to improve.
CONSOLIDATE	169	b145	My teacher takes the time to summarize what we learn each day.
	143	b70	In this class, we learn a lot almost every day.

Table A3. 3 Continued.

	164	b128	My teacher asks questions to be sure we are following along when s/he is teaching.
CHALLENGE		b133	My teacher asks students to explain more about answers they give.
	132	b21	In this class, my teacher accepts nothing less than our full effort.
		b36	My teacher doesn't let people give up when the work gets hard.
		b45	My teacher wants us to use our thinking skills, not just memorize things.
	110	b59	My teacher makes us explain our answers – why we think what we think.
CLASSROOM MANAGEMENT	138	b6	Our class stays busy and doesn't waste time.
		b49	Students in this class treat the teacher with respect.
	38	b46	My classmates behave the way my teacher wants them to.
		b138	Student behavior in this class is a problem.
		b114	Student behavior in this class makes the teacher angry.
		b113	I hate the way that students behave in this class.
		b112	Student behavior in this class is under control.

Table A3. 4 Detailed Fit Indices for Elementary School Data

Fit statistics	Value	Description
Likelihood ratio		
chi2_ms(13)	14.087	model vs. saturated
p>chi2	0.368	
chi2_bs(21)	175.201	baseline vs. saturated
p>chi2	0	
Population error		
RMSEA	0.015	Root mean squared error of approximation
90% CI, lower bound	0	
upper bound	0.056	
pclose	0.906	Probability RMSEA <=0.05
Information criteria		
AIC	4040.249	
BIC	4125.373	
Baseline comparison		
CFI	0.993	Comparative fit index
TLI	0.989	Tucker-Lewis index
Size of residuals		
SRMR	0.032	Standardized root mean squared residual
CD	0.597	Coefficient of determination

Table A3. 5 Detailed Fit Indices for Secondary School Data

Fit statistics	Value	Description
Likelihood ratio		
chi2_ms(13)	30.835	model vs. saturated
p>chi2	0.004	
chi2_bs(21)	181.869	baseline vs. saturated
p>chi2	0	
Population error		
RMSEA	0.072	Root mean squared error of approximation
90% CI, lower bound	0.039	
upper bound	0.105	
pclose	0.123	Probability RMSEA <=0.05
Information criteria		
AIC	3421.981	
BIC	3500.901	
Baseline comparison		
CFI	0.889	Comparative fit index
TLI	0.821	Tucker-Lewis index
Size of residuals		
SRMR	0.056	Standardized root mean squared residual
CD	0.781	Coefficient of determination

Table A3. 6 Analytic results for CKT impacts on CLASS and Tripod composite scores

VARIABLES	Regression Coefficients	Std. Errors	p-value
Elementary schools			
CLASS	.011	.376	.977
Tripod	-.765	.748	.306
Middle schools			
CLASS	.272	.389	.484
Tripod	-.042	1.604	.979
High schools			
CLASS	-.001	.881	.999
Tripod	3.764	3.314	.256

Standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Table A3. 7 Analytic results for dimension scores (outcomes centered by randomization block)

VARIABLES	Regression Coefficients	Std. Errors	p-value
Elementary schools			
InsQ	.167	.139	.229
MQI	.027*	.010	**.004**
CWCM	.021	.009	**.016**
Error and Imprecision	-.048	.020	**.015**
Richness of Mathematics	.055*	.019	**.003**
SPMMR	.016	.021	.436
WWSM	.053*	.018	**.003**
CLASS	.042	.284	.882
Tripod	-.704	.537	.190
Middle schools			
InsQ	.315	.227	.165
MQI	.021***	.006	**.000**
CWCM	-.001	.012	.940
Error and Imprecision	-.069***	.019	**.000**
Explicitness and Thoroughness	.096	.253	.840
Richness of Mathematics	.049*	.017	**.003**
SPMMR	.043*	.014	**.001**
WWSM	.024	.017	.162
CLASS	.266	.289	.359
Tripod	-.071	1.171	.952
High schools			
InsQ	.407	.404	.314
MQI	.007	.009	.428
Error and Imprecision	.010	.028	.711
Explicitness and Thoroughness	-.238	.477	.619
Richness of Mathematics	.042	.035	.230
SPMMR	-.016	.018	.395
WWSM	.011	.031	.713
CLASS	.289	.575	.615
Tripod	4.866	2.136	**.023**

Standard errors in parentheses

*** p<0.0001, ** p<0.001, * p<0.005 (p-value adjustment for 10 tests)

Table A3. 8 Analytic results for CKT impacts on Tripod 7Cs in high school classrooms (outcomes centered by randomization block)

	Regression Coefficients	Std. Errors	p-value
Care	.187	.075	**.013**
Confer	.120	.046	**.009**
Captivate	.254	.092	**.006**
Clarify	.187	.097	**.010**
Consolidate	.201	.076	**.008**
Challenge	.121	.055	**.027**
Classroom management	.137	.074	**.065**

Table A4. 1 Analytic Results for Step 3: Treatment Effects on Outcomes with Mediator and Treatment-by-mediator Interaction

	MQI	Tripod
Elementary schools (total effects: 0.017)		
Z, β_{0100}	.014	.012
	(.022)	(.022)
M, β_{0200}	.127	.0004
	(.142)	(.002)
ZM, β_{0300}	.175	.007
	(.344)	(.005)
Intercept, β_{0000}	.050	.051
	(.100)	(.100)
Student-level covariates	YES	YES
Grade indicators	YES	YES
District indicators	YES	YES
School	.133	.134
Random block	.023	.023
Class	.031	.031
Residual	.306	.306
Number of schools	74	74
Number of random blocks	109	109
Number of classrooms	267	267
Middle schools (total effect: 0.037)		
Z, β_{0100}	.025	.042*
	(.021)	(.021)
M, β_{0200}	.433	.002
	(.224)	(.001)
ZM, β_{0300}	-.819	-.008
	(.634)	(.022)
Intercept, β_{0000}	.337**	.315**
	(.109)	(.109)
Student-level covariates	YES	YES
Grade indicators	YES	YES
District indicators	YES	YES
School	.136	.141
Random block	.131	.128
Class	.022	.022

Table A4. 1 Continued.

Residual	.266	.266
Number of schools	74	74
Number of random blocks	114	114
Number of classrooms	238	238
<u>High schools (total effect: 0</u>.0002)		
Z, β_{0100}	.012	.006
	(.046)	(.040)
M, β_{0200}	-.890	.000004
	(.594)	(.003)
ZM, β_{0300}	-2.991**	-.006
	(1.109)	(.004)
Intercept, β_{0000}	-.662***	-.667***
	(.118)	(.138)
Student-level covariates	YES	YES
Grade indicators	YES	YES
District indicators	YES	YES
School	.012	.036
Random block	.000	.000
Class	.0001	.000
Residual	.240	.239
Number of schools	14	14
Number of random blocks	15	15
Number of classrooms	26	26

Appendix Figure

Figure A4. 1 Effective Instructional Practices Identified by Brophy and Good